Journal of the Royal Statistical Society Series B: Statistical Methodology, 2023, **85**, 1326–1354
https://doi.org/10.1093/jrsssb/qkad059
Advance access publication 10 June 2023
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Quasi-Newton updating for large-scale distributed learning

Shuyuan Wu[1], Danyang Huang[2] and Hansheng Wang[3]

[1]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China
[2]Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing, China
[3]Guanghua School of Management, Peking University, Beijing, China

Address for correspondence: Danyang Huang, Center for Applied Statistics and School of Statistics, Renmin University of China, 59 Zhongguancun Street, Beijing 100872, China. Email: dyhuang@ruc.edu.cn

Abstract

Distributed computing is critically important for modern statistical analysis. Herein, we develop a distributed quasi-Newton (DQN) framework with excellent statistical, computation, and communication efficiency. In the DQN method, no Hessian matrix inversion or communication is needed. This considerably reduces the computation and communication complexity of the proposed method. Notably, related existing methods only analyse numerical convergence and require a diverging number of iterations to converge. However, we investigate the statistical properties of the DQN method and theoretically demonstrate that the resulting estimator is statistically efficient over a small number of iterations under mild conditions. Extensive numerical analyses demonstrate the finite sample performance.

Keywords: communication efficiency, computation efficiency, distributed system, quasi-Newton methods, statistical efficiency

1 Introduction

Modern statistical analysis often involves massive datasets (Gopal & Yang, 2013). In several cases, such datasets are too large to be efficiently handled by a single computer. Instead, they have to be divided and then processed on a distributed computer system, which consists of a large number of computers (Zhang et al., 2013). Among all such computers, one often serves as the central computer, while the rest serve as worker computers. In this scenario, the central computer should be connected with all worker computers to construct a distributed computing system. Thus, approaches for the realisation of efficient statistical learning on such distributed computing systems have received considerable interest from the research community (Hector & Song, 2020, 2021; Jordan et al., 2019; Mcdonald et al., 2009; Tang et al., 2020).

Here, we consider a standard statistical learning problem with a total of N observations, where N is assumed to be very large. For each observation i, we collect a response variable $Y_i \in \mathbb{R}$ and corresponding feature vector $X_i \in \mathbb{R}^p$. The objective is to accurately estimate an unknown parameter $\theta_0 \in \mathbb{R}^p$ by minimising an appropriately defined empirical loss function (e.g. negative log-likelihood function), denoted by $\mathcal{L}(\theta) = \sum_{i=1}^{N} \ell(X_i, Y_i; \theta)$, where $\ell(X_i, Y_i; \theta)$ is the loss function defined on the ith sample. Under a traditional setup with a small sample size N, this optimisation problem can be easily solved using, for example, the standard Newton–Raphson algorithm. Specifically, let $\widehat{\theta}^{(0)}$ be an appropriate initial estimator of θ_0. Next, let $\widehat{\theta}^{(t)}$ be the estimator obtained in the tth iteration. Then, the $(t+1)$-th step estimator can be obtained as follows:

$$\widehat{\theta}^{(t+1)} = \widehat{\theta}^{(t)} - \alpha_t \{\ddot{\mathcal{L}}(\widehat{\theta}^{(t)})\}^{-1} \dot{\mathcal{L}}(\widehat{\theta}^{(t)}) \tag{1.1}$$

Received: October 25, 2022. **Revised:** March 20, 2023. **Accepted:** May 20, 2023

where $\dot{\mathcal{L}}(\theta)$ and $\ddot{\mathcal{L}}(\theta)$ represent the first- and second-order derivatives of the loss function $\mathcal{L}(\cdot)$ with respect to θ, respectively, and α_t represents the learning rate. Here, we assume that the initial estimator is close to θ_0, and thus, we set $\alpha_t = 1$ (Mokhtari et al., 2018). However, for a massive dataset that is distributed on a distributed computing system, efficient execution of the above Newton–Raphson algorithm becomes a nontrivial problem.

One straightforward solution is to retain the original Newton–Raphson algorithm but with distributed computing. Specifically, we assume that there exist M workers indexed by $1 \leq m \leq M$. We denote the entire sample by $\mathcal{S}_F = \{1, 2, \ldots, N\}$ and the sample allocated to the mth worker by $\mathcal{S}_{(m)} \subset \mathcal{S}_F$. Then, we have $\cup_{m=1}^{M} \mathcal{S}_{(m)} = \mathcal{S}_F$ and $\mathcal{S}_{(m_1)} \cap \mathcal{S}_{(m_2)} = \emptyset$ for any $m_1 \neq m_2$. Given $\widehat{\theta}^{(t)}$, we can then compute the first- and second-order derivatives of the loss function as follows:

$$\dot{\mathcal{L}}\big(\widehat{\theta}^{(t)}\big) = M^{-1} \sum_{m=1}^{M} \dot{\mathcal{L}}_{(m)}\big(\widehat{\theta}^{(t)}\big) \quad \text{and} \quad \ddot{\mathcal{L}}\big(\widehat{\theta}^{(t)}\big) = M^{-1} \sum_{m=1}^{M} \ddot{\mathcal{L}}_{(m)}\big(\widehat{\theta}^{(t)}\big)$$

where $\dot{\mathcal{L}}_{(m)}(\widehat{\theta}^{(t)}) = \sum_{i \in \mathcal{S}_{(m)}} \dot{\ell}(X_i, Y_i; \widehat{\theta}^{(t)})$ and $\ddot{\mathcal{L}}_{(m)}(\widehat{\theta}^{(t)}) = \sum_{i \in \mathcal{S}_{(m)}} \ddot{\ell}(X_i, Y_i; \widehat{\theta}^{(t)})$. $\dot{\ell}(X_i, Y_i; \widehat{\theta}^{(t)})$ and $\ddot{\ell}(X_i, Y_i; \widehat{\theta}^{(t)})$ are computed on the mth worker and are transferred to the central computer for updating $\widehat{\theta}^{(t+1)}$, according to (1.1). The solution is easy to implement and useful in practical applications but has several serious limitations. First, inverting the $p \times p$-dimensional Hessian matrix using the central computer incurs a computation cost on the order of $O(p^3)$ for each iteration. Second, transferring the local Hessian matrices from each worker to the central computer incurs a communication cost of order $O(p^2)$ for each worker in each iteration. Thus, this approach could incur high computation and communication costs for high-dimensional data (Fan et al., 2019).

Consequently, various communication-efficient Newton-type methods have been proposed to alleviate high communication costs. The underlying key idea is to maximally reduce the number of iterations required to transfer the Hessian matrix. For example, various one-step estimators have been proposed (Huang & Huo, 2019; F. Wang et al., 2020; Zhu et al., 2021). For these methods, only one round of Hessian matrix communication is needed. The resulting estimator can be statistically as efficient as the global one under appropriate regularity conditions. Methods avoiding Hessian matrix transmission have also been developed (Crane & Roosta, 2019; Jordan et al., 2019; Shamir et al., 2014; S. Wang et al., 2018; Zhang & Lin, 2015). The underlying key idea is to approximate the entire sample Hessian matrix using an appropriate local estimator, which is computed on a single computer (e.g. the central computer). Consequently, the communication cost resulting from Hessian transmission can be avoided. The inspiration for most statistical research on these methods is to obtain an estimator with statistical efficiency comparable to that of the global one within a small number of iterations. In this manner, the communication cost could be significantly reduced.

Nevertheless, the computation cost for calculating the Hessian inverse matrix is still of order $O(p^3)$. On one hand, to avoid matrix inverse calculation, distributed gradient descent algorithms have been developed (Goyal et al., 2017; Lin & Zhou, 2018; Qu & Li, 2019; Su & Xu, 2019), which require only first-order derivatives of the loss function (i.e. gradients). However, a large number of iterations are typically required for convergence, and the choice of hyperparameters is cumbersome. On the other hand, quasi-Newton methods in a distributed manner have been developed to address this problem (Chen et al., 2014; Eisen et al., 2017; Lee et al., 2018; Soori et al., 2020). The key idea behind quasi-Newton methods is to approximate the Hessian inverse in each iteration without actually inverting the matrix (Davidon, 1991; Goldfarb, 1970).

Specifically, for quasi-Newton methods, given an approximately inverted Hessian matrix in the tth iteration $H^{(t)}$, we could obtain $H^{(t+1)}$ by solving a linear equation, which is referred to as a secant condition (Davidon, 1991; Goldfarb, 1970):

$$H^{(t+1)}\big\{\dot{\mathcal{L}}(\widehat{\theta}^{(t+1)}) - \dot{\mathcal{L}}(\widehat{\theta}^{(t)})\big\} = (\widehat{\theta}^{(t+1)} - \widehat{\theta}^{(t)}) \tag{1.2}$$

Unfortunately, the secant condition cannot uniquely determine $H^{(t+1)}$. Two classical solutions have been proposed to solve this problem. The first is *symmetric rank one update* (Davidon, 1991; SR1). The second solution is referred to as *symmetric rank two update* (Goldfarb, 1970;

SR2), which is also called Broyden–Fletcher–Goldfarb–Shanno (BFGS) update. The distributed SR1 (Soori et al., 2020) and BFGS (Chen et al., 2014; Eisen et al., 2017) methods are correspondingly designed. The communication cost of these types of methods could have orders as low as $O(p)$ in each iteration. However, multiple rounds of communication are still required. Moreover, most of the existing studies discuss the numerical convergence of distributed quasi-Newton methods; however, discussions on statistical properties are limited.

To address this, we develop a novel distributed quasi-Newton (DQN) learning method that focuses on the statistical efficiency. With the help of a statistical discussion, we demonstrate that the proposed DQN algorithm requires only a small number of communication iterations to produce an estimator that is statistically as efficient as the global one. As a consequence, the proposed estimator is both communicationally and computationally efficient. Specifically, estimators and approximated Hessian inverses are first locally computed on each worker computer. Then, a communication mechanism is designed so that each worker passes the local Hessian information to the central computer but only in the form of a p-dimensional vector. In each iteration, the communication cost is of order $O(p)$, which is the same as that reported in most existing DQN-related studies (Chen et al., 2014; Eisen et al., 2017; Lee et al., 2018; Mokhtari et al., 2018; Soori et al., 2020). However, the proposed DQN method requires only a finite number of iterations with statistical guarantees. To be more precise, under the mild condition, i.e. $Np^{2K}(\log p)^{K+1}/n^{2K+2} \to 0$, only $3K$ rounds of iterations are required, where K is a small finite integer. Consequently, the overall costs attributed to communication and computation are statistically guaranteed. By contrast, a diverging number of iterations is required by methods presented in the existing literature.

The remainder of this paper is organised as follows. In Section 2, we present the DQN methodology and theoretical properties. Numerical studies, including simulation experiments and real data analysis, are presented in Section 3. Section 4 concludes the article with a brief discussion. All technical details are delegated to the appendixes.

2 Methodology

2.1 Quasi-Newton algorithm

We first introduce some notations for model definition. We consider a standard master-and-worker type distributed computation system with one central computer and M worker computers. Let us recall that $\mathcal{S}_{(m)}$ is the index set of the sample distributed to the mth worker. For convenience, we assume that $|\mathcal{S}_{(m)}| = n$ for every $1 \le m \le M$. Then, we have $N = nM$. Moreover, we recall that the global loss function is given by $\mathcal{L}(\theta) = N^{-1}\sum_{i=1}^{N} \ell(X_i, Y_i; \theta)$. We define $\widehat{\theta}_{\mathrm{ge}} = \mathrm{argmin}_\theta \mathcal{L}(\theta)$ and $\theta_0 = \mathrm{argmin}_\theta E\{\ell(X_i, Y_i; \theta)\}$ as the global estimator and true parameter, respectively. Under appropriate regularity conditions (Shao, 2003), we have $\sqrt{N}(\widehat{\theta}_{\mathrm{ge}} - \theta_0) \to_d N(0, \Sigma)$ for some positive definite matrix $\Sigma \in \mathbb{R}^{p\times p}$ as $N \to \infty$. For example, $\mathcal{L}(\theta)$ can be defined as twice the negative log-likelihood function. Accordingly, $\widehat{\theta}_{\mathrm{ge}}$ becomes the maximum likelihood estimator (MLE). Subsequently, we define the local loss function on the mth worker computer as $\mathcal{L}_{(m)}(\theta) = n^{-1}\sum_{i\in\mathcal{S}_m} \ell(X_i, Y_i; \theta)$. Let $\widehat{\theta}_{(m)} = \mathrm{argmin}_\theta \mathcal{L}_{(m)}(\theta)$ be the estimator locally obtained on the mth worker computer. Furthermore, $\dot{\ell}(X_i, Y_i; \theta) = \partial\ell(X_i, Y_i; \theta)/\partial\theta \in \mathbb{R}^p$, $\ddot{\ell}(X_i, Y_i; \theta) = \partial\ell(X_i, Y_i; \theta)/\partial\theta\theta^\top \in \mathbb{R}^{p\times p}$, and $\dddot{\ell}(X_i, Y_i; \theta) = \partial \mathrm{vec}\{\ddot{\ell}(X_i, Y_i; \theta)\}/\partial\theta \in \mathbb{R}^{p\times p^2}$ denote the first-, second-, and third-order derivatives of θ, respectively. Finally, for any matrix $B \in \mathbb{R}^{p\times q}$, $\|B\|_2$ is the maximum singular value of B. If B is a symmetric matrix, then $\lambda_{\min}(B)$ and $\lambda_{\max}(B)$ represent its minimal and maximal eigenvalues, respectively.

Before presenting the new method, we briefly introduce quasi-Newton methods. The well-known quasi-Newton methods were developed to address the problem of computation cost (Davidon, 1991; Goldfarb, 1970). The key idea is to approximate the Hessian inverse without inverting the Hessian matrix. Let $H^{(t)} \in \mathbb{R}^{p\times p}$ be the approximately inverted Hessian matrix used in the tth iteration. For the standard Newton–Raphson algorithm, we have $H^{(t)} = \{\ddot{\mathcal{L}}(\widehat{\theta}^{(t)})\}^{-1}$. However, for the quasi-Newton algorithm, this is defined in a different but smart manner. Specifically, note that $\dot{\mathcal{L}}(\widehat{\theta}^{(t+1)}) - \dot{\mathcal{L}}(\widehat{\theta}^{(t)}) \approx \ddot{\mathcal{L}}(\widehat{\theta}^{(t)})(\widehat{\theta}^{(t+1)} - \widehat{\theta}^{(t)})$ based on Taylor's expansion. This suggests that given $H^{(t)}$, we could obtain $H^{(t+1)}$ by solving the secant condition (1.2). As mentioned

previously, the secant condition cannot uniquely determine $H^{(t+1)}$. For SR1 update (Davidon, 1991), given $H^{(t)}$, $H^{(t+1)}$ is updated based on the rank one correction of $H^{(t)}$, i.e. $H^{(t+1)} = H^{(t)} + \alpha uu^\top$ for some undetermined coefficient $\alpha \in \mathbb{R}$ and $u \in \mathbb{R}^p$. Accordingly, solving α and u using (1.2), we obtain

$$H^{(t+1)} = H^{(t)} + \frac{v^{(t)}\{v^{(t)}\}^\top}{\{v^{(t)}\}^\top\{\dot{\mathcal{L}}(\widehat{\theta}^{(t+1)}) - \dot{\mathcal{L}}(\widehat{\theta}^{(t)})\}} \tag{2.1}$$

where $v^{(t)} = \widehat{\theta}^{(t+1)} - \widehat{\theta}^{(t)} - H^{(t)}\{\dot{\mathcal{L}}(\widehat{\theta}^{(t+1)}) - \dot{\mathcal{L}}(\widehat{\theta}^{(t)})\}$. For SR2 update (Goldfarb, 1970), given $H^{(t)}$, $H^{(t+1)}$ is updated according to

$$H^{(t+1)} = (V^{(t)})^\top H^{(t)} V^{(t)} + \rho^{(t)}\left(\widehat{\theta}^{(t+1)} - \widehat{\theta}^{(t)}\right)\left(\widehat{\theta}^{(t+1)} - \widehat{\theta}^{(t)}\right)^\top \tag{2.2}$$

where $\rho^{(t)} = 1/[(\widehat{\theta}^{(t+1)} - \widehat{\theta}^{(t)})^\top\{\dot{\mathcal{L}}(\widehat{\theta}^{(t+1)}) - \dot{\mathcal{L}}(\widehat{\theta}^{(t)})\}]$, $V^{(t)} = I_p - \rho^{(t)}\{\dot{\mathcal{L}}(\widehat{\theta}^{(t+1)}) - \dot{\mathcal{L}}(\widehat{\theta}^{(t)})\}(\widehat{\theta}^{(t+1)} - \widehat{\theta}^{(t)})^\top$, and $I_p \in \mathbb{R}^p$ is an identity matrix. Equation (2.2) is the well-known BFGS formula (Goldfarb, 1970). For convex loss functions, (2.2) guarantees the positive definiteness of $H^{(t+1)}$ if $H^{(t)}$ is positive definite (Nocedal & Wright, 1999). Derivation details for obtaining (2.1) and (2.2) are described in Appendix D.

Comparing (2.1) and (2.2) with $\{\ddot{\mathcal{L}}(\widehat{\theta}^{(t)})\}^{-1}$, we find that no Hessian matrix inversion is needed for computing $H^{(t+1)}$ using the SR1 or BFGS algorithm if the previous update $H^{(t)}$ is available. Thus, both algorithms offer highly efficient computation. After computing $H^{(t)}$, $\widehat{\theta}^{(t)}$ can be updated as $\widehat{\theta}^{(t+1)} = \widehat{\theta}^{(t)} - H^{(t)}\dot{\mathcal{L}}(\widehat{\theta}^{(t)})$. As proved in Broyden et al. (1973), the resulting estimator converges Q-superlinearly to the global estimator $\widehat{\theta}_{\text{ge}}$; i.e. $\|\widehat{\theta}^{(t+1)} - \widehat{\theta}_{\text{ge}}\|/\|\widehat{\theta}^{(t)} - \widehat{\theta}_{\text{ge}}\| \to 0$ as $t \to \infty$ for a strongly convex loss function. This convergence rate is slightly lower than the quadratic rate of the classical Newton–Raphson algorithm. However, it is much faster than the linear rate of various gradient-based methods. This makes the quasi-Newton algorithm one of the most popular algorithms in practice (Nocedal & Wright, 1999).

2.2 Distributed one-stage quasi-Newton estimator

To avoid multiple rounds of iterations, we consider the distributed one-stage quasi-Newton estimator. We were motivated to do so for two reasons. First, as mentioned previously, the quasi-Newton method is computationally efficient, because no Hessian matrix inversion is involved. This makes it particularly attractive for high-dimensional data analysis. Second, with an interesting modification, we find that the quasi-Newton algorithm can operate with a master-and-worker-type distributed computing system in a very natural and comfortable manner. The resulting communication cost is also minimal, that is, of order $O(p)$. Specifically, we present a communication and computation efficient distributed one-stage quasi-Newton algorithm, which can be executed using the following three steps.

Step 1. At the start, we assume that an initial estimator can be provided for the central computer. The initial estimator should be convenient to obtain. Moreover, it needs to be consistent, but excellent statistical efficiency is not necessary. Here, we consider the popularly used one-shot estimator (Zhang et al., 2013) as the initial estimator. Accordingly, we need each client to report a local estimator $\widehat{\theta}_{(m)}$ determined by the quasi-Newton algorithm to the central computer. Next, a global estimator can be simply assembled as $\widehat{\theta}_{\text{stage},0} = M^{-1}\sum_{m=1}^{M}\widehat{\theta}_{(m)}$. Once the initial estimator $\widehat{\theta}_{\text{stage},0}$ is obtained, it is then broadcasted to every worker computer; see the left panel of Figure 1. This completes the first round of communication with $O(p)$ cost.

Step 2. After receiving $\widehat{\theta}_{\text{stage},0}$ from the central computer, each worker computer can compute the local gradients $\dot{\mathcal{L}}_{(m)}(\widehat{\theta}_{\text{stage},0})$. These are then transferred back to the central computer to determine the global gradient $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0}) = M^{-1}\sum_{m=1}^{M}\dot{\mathcal{L}}_{(m)}(\widehat{\theta}_{\text{stage},0})$. Thereafter, the global gradient $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})$ should be broadcasted back to each worker computer. The middle panel of Figure 1

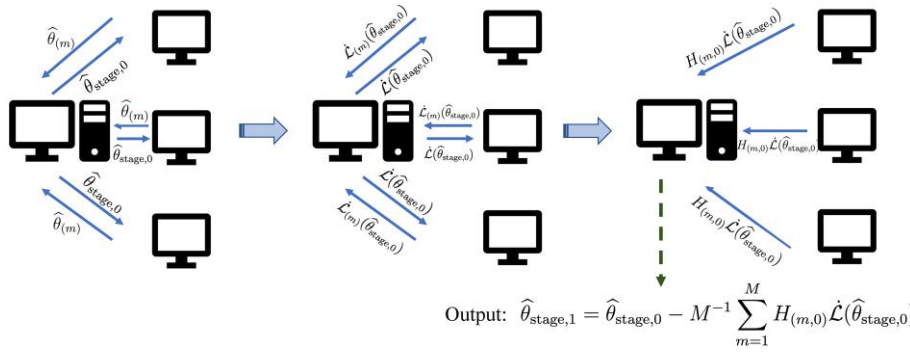

Figure 1. Illustration of the communication-efficient one-stage method.

presents an illustration of this second step. This completes the second round of communication with $O(p)$ cost.

Step 3. When deriving $\widehat{\theta}_{(m)}$ in Step 1, the approximated Hessian inverse $H_{(m,0)}$ is also obtained as a byproduct of the quasi-Newton algorithm. We apply $H_{(m,0)}$ to the global gradient vector $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})$ to obtain a p-dimensional vector $H_{(m,0)}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})$, which is then reported back to the central computer. This completes the third round of communication with $O(p)$ cost. Recall that the central computer also holds the initial estimator $\widehat{\theta}_{\text{stage},0}$. Then, a new estimator can be obtained as follows:

$$\widehat{\theta}_{\text{stage},1} = \widehat{\theta}_{\text{stage},0} - M^{-1}\sum_{m=1}^{M} H_{(m,0)}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})$$

The right panel of Figure 1 presents an illustration of the last step. For convenience, we refer to $\widehat{\theta}_{\text{stage},1}$ as the one-stage DQN estimator (DQN(1)).

To summarise, three rounds of master-and-worker communication are needed to compute $\widehat{\theta}_{\text{stage},1}$. Because the communication cost for each round is of order $O(p)$, the total communication cost is also of the same order, which is the lowest communication complexity possible for a p-dimensional distributed parameter estimation problem. A more detailed description of the algorithm is given in Algorithm 1. Note that $\widehat{\theta}_{\text{stage},1}$ shares a similar spirit as the classical one-step estimator for MLE (Van der Vaart, 2000). However, $\widehat{\theta}_{\text{stage},1}$ is mainly designed for a distributed system with minimal communication and computation costs.

2.3 Theoretical properties

We next study the theoretical properties of the proposed DQN(1) estimator. To this end, several regularity conditions must be considered.

(C1) (Randomness) Assume that (X_i, Y_i)s on the mth worker are independently and identically distributed.

(C2) (Parameters) The parameter space $\boldsymbol{\Theta}$ is a compact and convex subset of $\mathbb{R}^p$. In addition, $\theta_0 \in \text{int}(\boldsymbol{\Theta})$ and $R := \sup_{\theta\in\boldsymbol{\Theta}} \|\theta - \theta_0\| > 0$.

(C3) (Local strong convexity) Define $\Omega(\theta) = E[\dot{\ell}(X_i, Y_i; \theta)\{\dot{\ell}(X_i, Y_i; \theta)\}^{\top}] = -E\{\ddot{\ell}(X_i, Y_i; \theta)\}$. Assume $\tau_{\min} \le \lambda_{\min}\{\Omega(\theta_0)\} \le \lambda_{\max}\{\Omega(\theta_0)\} \le \tau_{\max}$ for some positive constants $\tau_{\min}$ and $\tau_{\max}$.

(C4) (Smoothness) Define $B(\theta_0, \delta) = \{\theta \in \mathbb{R}^p | \|\theta - \theta_0\| \le \delta\}$ to be a ball around θ_0 with radius $\delta > 0$. Assume that there exist two constants $C_G > 0$ and $C_H > 0$ such that the following inequalities hold.

$$E\left\{\left\|\dot{\ell}(X_i, Y_i; \theta)\right\|_2^8\right\} \le C_G^8, E\left\{\left\|\ddot{\ell}(X_i, Y_i; \theta) - \Omega(\theta)\right\|_2^8\right\} \le C_H^8 \quad \text{for all } \theta \in B(\theta_0, \delta).$$

Algorithm 1 Distributed one-stage quasi-Newton algorithm

Input: Initial estimators $\widehat{\theta}_{(m)}^{(0)}$, $H_{(m,0)}^{(0)}$ on the m-th worker, tolerance $\delta > 0$, maximum iterations T
Output: One-stage estimator $\widehat{\theta}_{\text{stage},1}$
for $m = 1, 2, \ldots M$ (distributedly) **do**
  While tol $> \delta$ and $t < T$ **do**
    $\widehat{\theta}_{(m)}^{(t+1)} = \widehat{\theta}_{(m)}^{(t)} - H_{(m,0)}^{(t)} \dot{\mathcal{L}}_{(m)}(\widehat{\theta}_{(m)}^{(t)})$, where $H_{(m,0)}^{(t)}$ is updated by (2.1) or (2.2)
  end
  Save $H_{(m,0)}^{(t)}$ and $\widehat{\theta}_{(m)}^{(t)}$ at convergence as $H_{(m,0)}$ and $\widehat{\theta}_{(m)}$ and then transfer $\widehat{\theta}_{(m)}$ to the central computer
end
The central computer computes $\widehat{\theta}_{\text{stage},0} = M^{-1} \sum_{m=1}^{M} \widehat{\theta}_{(m)}$ and broadcasts $\widehat{\theta}_{\text{stage},0}$ to each worker
for $m = 1, 2, \ldots M$ (distributedly) **do**
  Compute $\dot{\mathcal{L}}_{(m)}(\widehat{\theta}_{\text{stage},0})$ and transfer it to the central computer
end
The central computer computes $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0}) = M^{-1} \sum_{m=1}^{M} \dot{\mathcal{L}}_{(m)}(\widehat{\theta}_{\text{stage},0})$ and broadcasts $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})$ to each worker
for $m = 1, 2, \ldots M$ (distributedly) **do**
  Calculate $H_{(m,0)} \dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})$ and transfer it to the central computer
end
The central computer computes $\widehat{\theta}_{\text{stage},1} = \widehat{\theta}_{\text{stage},0} - M^{-1} \sum_{m=1}^{M} H_{(m,0)} \dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})$.

Moreover, for all $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$, $\ddot{\ell}(X_i, Y_i; \theta)$ and $\ddot{\ell}\,(X_i, Y_i; \theta)$ are both Lipschitz continuous, in the sense that for any $\theta', \theta'' \in B(\theta_0, \delta)$ and $u \in \mathbb{R}^p$,

$$\left\| \ddot{\ell}(X_i, Y_i; \theta') - \ddot{\ell}(X_i, Y_i; \theta'') \right\|_2 \leq C(X_i, Y_i) \left\| \theta' - \theta'' \right\| \text{ and}$$

$$\left\| \left\{ \ddot{\ell}\,(X_i, Y_i; \theta') - \ddot{\ell}\,(X_i, Y_i; \theta'') \right\} (u \otimes I_p) \right\|_2 \leq C(X_i, Y_i) \left\| \theta' - \theta'' \right\| \|u\|$$

and $E\{C^8(X_i, Y_i)\} \leq C_{\max}^8$, $E[C^8(X_i, Y_i) - E\{C(X_i, Y_i)\}^8] \leq C_{\max}^8$ for some positive constant $C_{\max}$.

(C5) (Convergence) For the mth worker, define the tth step local approximate Hessian inverse to be $H_{(m,0)}^{(t)}$, assume that $\lim_{t \to \infty} \|H_{(m,0)}^{(t)} - \{\ddot{\mathcal{L}}(\widehat{\theta}_{(m)})\}^{-1}\|_2 \to 0$.

(C6) (Dimensionality) We assume that $p\sqrt{\log p}/n \to 0$ as $n \to \infty$.

Condition (C1) requires that the data be randomly distributed across different computers to ensure the statistical consistency of the one-shot estimator as a convenient initial estimator. The same condition was adopted in Zhang et al. (2013) and Fan et al. (2019). (C2)–(C4) are classical regularity conditions in convex optimisation (Jordan et al., 2019; Zhang et al., 2013). (C5) guarantees the convergence of the approximation matrix $H_{(m,0)}$ and has previously been rigorously investigated. Specifically, for SR1 update, (C5) has been rigorously proved by Conn et al. (1991), assuming that sequence $\{H_{(m,0)}^{(t)} \dot{\mathcal{L}}(\widehat{\theta}^{(t)})\}$ is uniformly linearly independent. (C5) has also been verified for BFGS update by Schuller (1974) under slightly stronger conditions. (C6) specifies the relationship between the dimension p and local data size n. Given these technical conditions, we establish Theorem 1.

Theorem 1 Assume that (C1)–(C6) hold. Then, we have $\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\| \leq \kappa(M^{-1} \sum_{m=1}^{M} [\|\widehat{\theta}_{(m)} - \theta_0\|^2 + \|\ddot{\mathcal{L}}_{(m)}(\theta_0) - \Omega(\theta_0)\|_2^2 + \|\{\ddot{\mathcal{L}}_{(m)}(\theta_0) - \dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)} - \theta_0) \otimes I_p\}\|_2] + \|\widehat{\theta}_{\text{stage},0} - \theta_0\|)\|\widehat{\theta}_{\text{stage},0} - \theta_0\|$ for some constant $\kappa > 0$ with probability tending to one. Furthermore, assuming that $N(p \log p)^2/n^4 \to 0$, we have $\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\| = o_p(N^{-1/2})$.

The detailed proof is given in Appendix A.1. From Theorem 1, we infer that the discrepancy between $\widehat{\theta}_{\text{stage},1}$ and $\widehat{\theta}_{\text{ge}}$ is upper bounded by $(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)}-\theta_0\|^2+\|\ddot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2+\|\{\ddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}\|_2]+\|\widehat{\theta}_{\text{stage},0}-\theta_0\|)\|\widehat{\theta}_{\text{stage},0}-\theta_0\|=O_p(p\log p/n^2)+o_p(1/\sqrt{N})$. Thus, the difference $\|\widehat{\theta}_{\text{stage},1}-\widehat{\theta}_{\text{ge}}\|$ is further reduced compared with $\|\widehat{\theta}_{\text{stage},0}-\theta_0\|$ to order $O_p(1/\sqrt{N}+\sqrt{\log p}/n)$; see details in Equation (A.4) of Appendix A. The amount of compression is determined by three factors: (1) averaged distance of the local estimator $M^{-1}\sum_{m=1}^{M}\|\widehat{\theta}_{(m)}-\theta_0\|^2$; (2) averaged distance of the local estimator, Hessian matrix, and third derivative matrix $M^{-1}\sum_{m=1}^{M}[\|\ddot{\mathcal{L}}_{(m)}(\theta)-\Omega(\theta_0)\|_2^2+\|\{\ddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}\|_2]$; and (3) distance between the initial estimator and true parameter $\|\widehat{\theta}_{\text{stage},0}-\theta_0\|$. Accordingly, assuming $N(p\log p)^2/n^4\to 0$, $\widehat{\theta}_{\text{stage},1}$ achieves the optimal statistical efficiency. When p is fixed, this condition reduces to $N/n^4\to 0$. It is a condition much weaker than $N/n^2\to 0$, which has been typically assumed in the existing literature (F. Wang et al., 2020; Zhang et al., 2013).

2.4 Distributed multi-stage quasi-Newton estimator

In the previous section, we introduced the DQN(1) estimator. Note that to achieve the optimal statistical efficiency, we require $N(\log p)^4/n^4\to 0$. This condition can be easily satisfied if the feature dimension p is not too high. By contrast, if p is relatively high, the convergence rate of the DQN(1) estimator slows dow n. To fix this, we further develop a multi-stage DQN estimator. First, we present a two-stage DQN estimator with two extra updating steps with BFGS update and refer to it as the DQN(2) estimator. The details of the DQN(2) algorithm are given below. It is remarkable that, after the first three steps, the DQN(1) estimator $\widehat{\theta}_{\text{stage},1}$ is already computed by the central computer.

Step 4. Broadcasting the DQN(1) estimator to each worker computer. Similar to the DQN(1) algorithm, the worker computer should compute the local gradient $\dot{\mathcal{L}}_{(m)}(\widehat{\theta}_{\text{stage},1})$, which should be reported back to the central computer. As a consequence, the global gradient $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})$ can be assembled. This leads to two rounds of communication with a cost of order $O(p)$.

Step 5. Note that, when we compute the DQN(1) algorithm, each worker holds an approximated Hessian inverse matrix $H_{(m,0)}$. Moreover, note that $\widehat{\theta}_{\text{stage},0}$ and $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})$ are the estimators obtained in the process of the DQN(1) algorithm for each worker. Consequently, given $H_{(m,0)}$, each worker could compute the updated matrix $H_{(m,1)}$ according to the BFGS formula (2.2) as follows:

$$H_{(m,1)}=\{V_0\}^{\top}H_{(m,0)}\{V_0\}+\rho_0\left(\widehat{\theta}_{\text{stage},1}-\widehat{\theta}_{\text{stage},0}\right)\left(\widehat{\theta}_{\text{stage},1}-\widehat{\theta}_{\text{stage},0}\right)^{\top} \tag{2.3}$$

where $V_0=I_p-\rho_0\{\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})-\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})\}(\widehat{\theta}_{\text{stage},1}-\widehat{\theta}_{\text{stage},0})^{\top}$ and $\rho_0=1/[(\widehat{\theta}_{\text{stage},1}-\widehat{\theta}_{\text{stage},0})^{\top}\{\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})-\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})\}]$. After computing $H_{(m,1)}$, it is applied to the global gradient $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})$. This leads to a p-dimensional vector $H_{(m,1)}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})$, which is then reported back to the central computer. Subsequently, the DQN(2) estimator could be derived as

$$\widehat{\theta}_{\text{stage},2}=\widehat{\theta}_{\text{stage},1}-M^{-1}\sum_{m=1}^{M}H_{(m,1)}\dot{\mathcal{L}}\left(\widehat{\theta}_{\text{stage},1}\right) \tag{2.4}$$

Thus, Steps 4 and 5 constitute the second stage estimation. The detailed algorithm is given in Algorithm 2. Moreover, a two-stage DQN estimator with the SR1 updating strategy could be similarly obtained; more details are presented in Appendix C.1.

Similar to Algorithm 1, Algorithm 2 incurs another three rounds of communication. Recall that three extra rounds of communication are needed for computing $\widehat{\theta}_{\text{stage},1}$. Thus, a total of six rounds of communication are needed for computing $\widehat{\theta}_{\text{stage},2}$. The communication cost for each round remains of order $O(p)$. Consequently, the total communication cost of the two-stage estimator remains of order $O(p)$. Additionally, no Hessian matrix needs to be inverted. However, a better estimation accuracy could be achieved due to the additional updating stage, which leads to the next theorem.

Algorithm 2 Distributed two-stage quasi-Newton algorithm

Input: DQN(1) estimator $\widehat{\theta}_{\text{stage},1}$ on the central computer, $\widehat{\theta}_{\text{stage},0}$, $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})$, and the initial Hessian inverse approximation $H_{(m,0)}$ on the m-th worker

Output: DQN(2) estimator $\widehat{\theta}_{\text{stage},2}$

The central computer broadcasts $\widehat{\theta}_{\text{stage},1}$ to each worker

for $m = 1, 2, \ldots, M$ (distributedly) **do**

  Compute $\dot{\mathcal{L}}_{(m)}(\widehat{\theta}_{\text{stage},1})$ and transfer it to the central computer

end

The central computer computes $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) = M^{-1}\sum_{m=1}^{M}\dot{\mathcal{L}}_{(m)}(\widehat{\theta}_{\text{stage},1})$ and broadcasts $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})$ to each worker

for $m = 1, 2, \ldots, M$ (distributedly) **do**

  Update $H_{(m,1)}$ according to (2.3).

  Calculate $H_{(m,1)}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})$ and transfer it to the central computer

end

The central computer computes $\widehat{\theta}_{\text{stage},2} = \widehat{\theta}_{\text{stage},1} - M^{-1}\sum_{m=1}^{M} H_{(m,1)}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})$.

Theorem 2 Assume that the technical conditions (C1)–(C6) hold. Then, we have $\|\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{ge}}\| \le \kappa_2(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)} - \theta_0\|^2 + \|\ddot{\mathcal{L}}_{(m)}(\theta_0) - \Omega(\theta_0)\|_2^2 + \|\{\ddot{\mathcal{L}}_{(m)}(\theta_0) - \dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)} - \theta_0)\otimes I_p\}\|_2] + \|\widehat{\theta}_{\text{stage},0} - \theta_0\|)\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|$ for some constant $\kappa_2 > 0$ with probability tending to one. Furthermore, assuming that $Np^4(\log p)^3/n^6 \to 0$, we have $\|\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{ge}}\| = o_p(N^{-1/2})$.

The proof of Theorem 2 is given in Appendix A.2. It could be verified that the discrepancy between $\widehat{\theta}_{\text{stage},2}$ and $\widehat{\theta}_{\text{ge}}$ is further reduced from $\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\| = O_p(p\log p/n^2) + o_p(1/\sqrt{N})$ to $\|\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{ge}}\| = O_p(p^2(\log p)^{3/2}/n^3) + o_p(1/\sqrt{N})$; see Appendix A for more details. Accordingly, the optimal statistical efficiency can be achieved if $Np^4(\log p)^3/n^6 \to 0$. This is a weaker condition than that of the DQN(1) estimator. Next, we extend the idea of the DQN(2) estimator to develop the multi-stage DQN (DQN(K)) estimator $\widehat{\theta}_{\text{stage},K}$. The detailed algorithm is given in Algorithm 3. The theoretical properties are summarised by Corollary 1.

Corollary 1 Assume that the technical conditions (C1)–(C6) hold. Then, we have $\|\widehat{\theta}_{\text{stage},K} - \widehat{\theta}_{\text{ge}}\| \le \kappa_K(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)} - \theta_0\|^2 + \|\ddot{\mathcal{L}}_{(m)}(\theta_0) - \Omega(\theta_0)\|_2^2 + \|\{\ddot{\mathcal{L}}_{(m)}(\theta_0) - \dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)} - \theta_0)\otimes I_p\}\|_2] + \|\widehat{\theta}_{\text{stage},0} - \theta_0\|)^K\|\widehat{\theta}_{\text{stage},0} - \theta_0\|$ for some cnstant $\kappa_K > 0$ with probability tending to one. Furthermore, assuming that $Np^{2K}(\log p)^{K+1}/n^{2K+2} \to 0$, we have $\|\widehat{\theta}_{\text{stage},K} - \widehat{\theta}_{\text{ge}}\| = o_p(N^{-1/2})$.

The proof of Corollary 1 is given in Appendix A.3. It could be found that Corollary 1 is a directly generalised version of Theorem 2. To be more specific, the discrepancy between the DQN(K) estimator $\widehat{\theta}_{\text{stage},K}$ and $\widehat{\theta}_{\text{ge}}$ is further compressed from $\|\widehat{\theta}_{\text{stage},0} - \widehat{\theta}_{\text{ge}}\|$ by $(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)} - \theta_0\|^2 + \|\ddot{\mathcal{L}}_{(m)}(\theta_0) - \Omega(\theta_0)\|_2^2 + \|\{\ddot{\mathcal{L}}_{(m)}(\theta_0) - \dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)} - \theta_0)\otimes I_p\}\|_2] + \|\widehat{\theta}_{\text{stage},0} - \theta_0\|)^K$. Consequently, the optimal statistical efficiency can be obtained using the DQN(K) estimator with even weaker technical conditions. In other words, $Np^{2K}(\log p)^{K+1}/n^{2K+2} \to 0$. Moreover, Algorithm 3 shows that the DQN(K) estimator requires $3K$ rounds of communication, with cost $O(p)$ for each round. Therefore, practical applications should consider the trade-off between statistical efficiency and time cost.

3 Numerical studies

3.1 Performance of the DQN algorithm

We start with demonstrating the finite sample performance of the proposed DQN method. Specifically, we present two simulation examples as follows.

Algorithm 3 Distributed Distributed K-stage quasi-Newton algorithm

Input: DQN($K-1$) estimator $\widehat{\theta}_{\text{stage},K-1}$ on the central computer, $\widehat{\theta}_{\text{stage},K-2}$, $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},K-2})$, and Hessian inverse approximation $H_{(m,K-2)}$ on the m-th worker

Output: DQN(K) estimator $\widehat{\theta}_{\text{stage},K}$

The central computer broadcasts $\widehat{\theta}_{\text{stage},K-1}$ to each worker

for $m = 1, 2, \ldots, M$ (distributedly)**do**

  Compute $\dot{\mathcal{L}}_{(m)}(\widehat{\theta}_{\text{stage},K-1})$ and transfer it to the central computer

end

The central computer computes $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},K-1}) = M^{-1}\sum_{m=1}^{M}\dot{\mathcal{L}}_{(m)}(\widehat{\theta}_{\text{stage},K-1})$ and broadcasts $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},K-1})$ to each worker

for $m = 1, 2, \ldots, M$ (distributedly) **do**

  Compute $H_{(m,K-1)} = \{V_{K-2}\}^{\top} H_{(m,K-2)}\{V_{K-2}\} + \rho_{K-2}(\widehat{\theta}_{\text{stage},K-1} - \widehat{\theta}_{\text{stage},K-2})(\widehat{\theta}_{\text{stage},K-1} - \widehat{\theta}_{\text{stage},K-2})^{\top}$

  Calculate $H_{(m,K-1)}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})$ and transfer it to the central computer

end

The central computer obtains

$\widehat{\theta}_{\text{stage},K} = \widehat{\theta}_{\text{stage},K-1} - M^{-1}\sum_{m=1}^{M} H_{(m,K-1)}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},K-1})$.

- Example 1 (Logistic regression). We consider a logistic regression, which is one of the most popular classification models. We set $\theta_0 = c_0\gamma/\|\gamma\|$, where $\gamma \in \mathbb{R}^p$ is generated from a standard normal distribution, and $c_0 = 1.5$ controls the signal strength. The covariate X_i is generated from a multivariate normal distribution with $E(X_i) = 0$ and $\text{cov}(X_{ij_1}, X_{ij_2}) = \rho^{|j_1 - j_2|}$ with $\rho = 0.5$ for $1 \le j_1, j_2 \le p$. Given X_i, the response $Y_i \in \{0, 1\}$ is then generated according to $P(Y_i = 1|X_i, \theta_0) = \{1 + \exp(-X_i^{\top}\theta_0)\}^{-1}$.
- Example 2 (Poisson regression). This is an example revised from Fan and Li (2001). Specifically, θ_0 and X_i are the same as those in Example 1 but with $c_0 = 0.3$ and $\rho = 0.2$. Conditional on X_i, response Y_i is generated from a Poisson distribution with $E(Y_i|X_i) = \exp(X_i^{\top}\theta_0)$.

For each simulation example, the sample size is $N = 10^6$, and we vary the local data size n and dimension p. The generated sample data are randomly distributed to different workers $M = N/n$. We replicate the experiment $R = 100$ times for reliability.

To gauge the finite sample performance of the proposed method, various performance analyses are developed. Specifically, let $\widehat{\theta}^{(r)}_{\text{stage},K}$ be the DQN(K) estimator (by SR1 or BFGS updating) obtained in the rth replication. Then, the mean squared error (MSE) is defined as $\text{MSE} = R^{-1}\sum_{r=1}^{R}\|\widehat{\theta}^{(r)}_{\text{stage},K} - \theta_0\|^2$. Moreover, a total of four measures are developed to evaluate the estimator's stability and robustness. They are the MSE values in log-scale (i.e. log(MSE)), standard deviation (SD) of log(MSE), inter-quartile range (IQR) of log(MSE), and range of log(MSE). The detailed results are given in Tables 1 and 2. Because simulation results of Example 1 are quantitatively similar to those of Example 2, we report the results for Example 1 only. The detailed results for Example 2 are given in Appendix E.

From Table 1, we find that the values of all four measures increase as p decreases for a fixed n. By contrast, from Table 2, we find that, with a fixed p, larger n always leads to an improved estimation performance in the sense that all four measure values approach those of MLE (i.e. $\widehat{\theta}_{\text{ge}}$). Moreover, when p is relatively small or n is relatively large, the log(MSE) value of $\widehat{\theta}_{\text{stage},1}$ or $\widehat{\theta}_{\text{stage},2}$ is comparable with that of MLE. However, as p grows (or n drops), more stages (i.e. larger K) are required to obtain an estimator with optimal statistical efficiency. Nevertheless, the number of required stages K remains very small (e.g. $K \le 4$). Thus, the algorithm is communicationally and computationally efficient. The SD, IQR, and range values also demonstrate similar patterns. These results are consistent with our theoretical findings in Theorems 1 and 2 and Corollary 1.

Table 1. Log(MSE) values and the corresponding SD, IQR, and range values for Example 1

		Stage 0		Stage 1		Stage 2		Stage 3		Stage 4		
	p	SR1	BFGS	SR1	BFGS	SR1	BFGS	SR1	BFGS	SR1	BFGS	MLE
Log	1	−6.86	−6.86	−6.96	−6.96	−6.96	−6.96	−6.96	−6.96	−6.96	−6.96	−6.96
(MSE)	10	−3.91	−3.91	−4.61	−4.61	−4.62	−4.62	−4.64	−4.64	−4.65	−4.65	−4.65
	20	−2.65	−2.65	−3.87	−3.87	−3.90	−3.90	−3.93	−3.93	−3.96	−3.96	−3.96
SD	1	0.15	0.15	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16
	10	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
	20	0.03	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
IQR	1	0.19	0.19	0.19	0.19	0.20	0.20	0.21	0.21	0.20	0.20	0.19
	10	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06
	20	0.04	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Range	1	0.68	0.68	0.69	0.69	0.69	0.69	0.70	0.70	0.70	0.70	0.70
	10	0.26	0.26	0.23	0.23	0.23	0.23	0.32	0.32	0.23	0.23	0.23
	20	0.18	0.18	0.20	0.20	0.20	0.20	0.19	0.19	0.19	0.19	0.20

Note. The numerical performance is evaluated for different methods with different feature dimensions $p(\times 10^2)$. The whole sample size N and local sample size n are fixed to be 10^6 and 2×10^4, respectively. The reported results are averaged for $R = 100$ simulation replications.
MSE = mean squared error; SD = standard deviation; IQR = inter-quartile range; BFGS = Broyden–Fletcher–Goldfarb–Shanno; MLE = maximum likelihood estimator.

Table 2. Log(MSE) values and the corresponding SD, IQR, and range values for Example 1

		Stage 0		Stage 1		Stage 2		Stage 3		Stage 4		
	n	SR1	BFGS	SR1	BFGS	SR1	BFGS	SR1	BFGS	SR1	BFGS	MLE
Log	50	−2.97	−2.97	−5.23	−5.23	−5.12	−5.12	−5.29	−5.29	−5.32	−5.32	−5.34
(MSE)	100	−4.23	−4.23	−5.28	−5.28	−5.30	−5.30	−5.32	−5.32	−5.33	−5.33	−5.34
	500	−5.25	−5.25	−5.33	−5.33	−5.33	−5.33	−5.33	−5.33	−5.33	−5.33	−5.34
SD	50	0.04	0.04	0.08	0.08	0.29	0.29	0.08	0.08	0.08	0.08	0.08
	100	0.05	0.05	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
	500	0.07	0.07	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
IQR	50	0.04	0.04	0.10	0.10	0.18	0.18	0.10	0.10	0.10	0.10	0.09
	100	0.07	0.07	0.10	0.10	0.10	0.10	0.09	0.09	0.09	0.09	0.09
	500	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09
Range	50	0.22	0.22	0.40	0.40	1.33	1.33	0.41	0.41	0.42	0.42	0.43
	100	0.31	0.31	0.40	0.40	0.40	0.40	0.41	0.41	0.42	0.42	0.43
	500	0.42	0.42	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43

Note. The numerical performance is evaluated for different $n(\times 10^2)$ and methods. The whole sample size N and feature dimension p are fixed to $N = 10^6$ and $p = 10^3$, respectively. Finally, the reported results are averaged based on $R = 100$ simulations.
MSE = mean squared error; SD = standard deviation; IQR = inter-quartile range; BFGS = Broyden–Fletcher–Goldfarb–Shanno; MLE = maximum likelihood estimator.

3.2 Comparison with competing methods

We next compare the proposed method with the following four competing methods: (1) the distributed one-step Newton (DOSN) estimator of Huang and Huo (2019), (2) the communication-efficient surrogate likelihood (CSL) based estimator of Jordan et al. (2019), (3)

Table 3. Averaged computation cost T_1, communication cost T_2, and total time cost T for Examples 1 and 2

	p	DOSN	CSL	DMGD	DAQN	DQN-BFGS	DQN-SR1
				Example 1			
T_1	500	0.95	1.86	40.29	3.94	0.59	0.68
	1,000	2.71	5.54	70.49	13.44	1.11	1.36
	2,000	8.69	24.75	113.59	33.92	3.63	3.72
	2,500	13.14	45.37	267.96	68.14	5.57	7.27
T_2	500	42.16	0.95	21.36	1.56	0.85	0.67
	1,000	212.94	1.90	42.82	3.01	1.66	1.26
	2,000	799.87	3.86	72.77	5.56	3.46	2.42
	2,500	1,240.81	4.64	90.97	8.36	4.61	3.16
T	500	43.12	2.82	61.65	5.51	1.44	1.35
	1,000	215.64	7.44	113.32	16.45	2.76	2.62
	2,000	808.56	28.61	186.35	39.48	7.09	6.15
	2,500	1,253.95	50.01	358.93	76.51	10.17	10.42
				Example 2			
T_1	500	0.50	0.72	52.09	2.44	0.31	0.31
	1,000	1.67	2.24	70.20	7.56	0.56	0.60
	2,000	8.26	10.26	113.54	26.62	1.50	1.53
	2,500	13.22	18.94	144.30	53.25	1.95	2.27
T_2	500	40.47	1.07	19.09	1.00	0.36	0.31
	1,000	166.58	1.68	41.57	2.20	0.69	0.56
	2,000	775.31	3.09	72.49	5.35	1.42	1.18
	2,500	1,308.45	3.72	91.92	6.79	1.80	1.58
T	500	40.96	1.79	71.18	3.44	0.67	0.62
	1,000	168.24	3.92	111.77	9.76	1.25	1.16
	2,000	783.57	13.35	186.03	31.96	2.92	2.71
	2,500	1,321.68	22.66	236.23	60.04	3.75	3.85

Note. The time cost is evaluated for different methods with different feature dimensions p. The whole sample size N and local sample size n are fixed to be 10^6 and 2×10^4, respectively. The reported results are averaged for $R = 100$ simulation replications.
DOSN = distributed one-step Newton; CSL = communication-efficient surrogate likelihood; DMGD = distributed momentum gradient descent; DAQN = distributed asynchronous averaged quasi-Newton; DQN = distributed quasi-Newton; BFGS = Broyden–Fletcher–GoldfarbShanno

the distributed momentum gradient descent (DMGD) estimator of Goyal et al. (2017), and (4) the distributed asynchronous averaged quasi-Newton (DAQN) estimator of Soori et al. (2020). The simulation model used here is the same as that in Section 3.1. We fix the sample size to be $N = 10^6$, the number of workers to be $M = 50$, and vary the dimension p from 500 to 2,500. Moreover, we set $K = 4$ for Example 1 and $K = 2$ for Example 2. We replicate the experiment for a total of $R = 100$ times.

To gauge the finite performance of different methods, we consider four different performance measures. First, to measure the *estimation accuracy*, we focus on the log(MSE) values. Second, to compare *computation efficiency*, we record for each method the computing time for the master plus the averaged computing time for each worker as $T_1^{(r)}$ in the rth ($1 \le r \le R$) simulation. Then, the averaged computing time T_1 for the R simulations is calculated and reported. Third, to measure the *communication efficiency*, the communication time for each simulation $T_2^{(r)}$ is estimated by the overall time cost $T^{(r)}$ minus the computing time $T_1^{(r)}$. Similarly, the averaged communication time T_2 is calculated and reported. Finally, the averaged *total time cost* T is also reported for better comparison. The simulation results are reported in Table 3 and Figure 2.

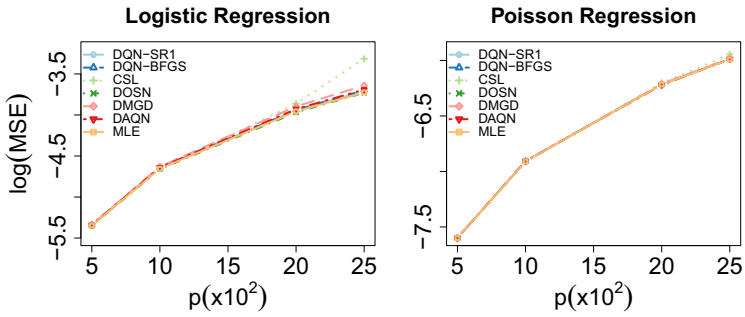

Figure 2. Log(MSE) values for different methods with different dimension p. The results for the logistic and Poisson regression models are given in the left and right panels, respectively. The whole sample size N and local subsample size n are fixed at $N = 10^6$ and $n = 2 \times 10^4$. The number of workers $M = 50$. The reported log(MSE) values are averages for $R = 100$ simulations. MSE = mean squared error.

As shown in Figure 2, all methods demonstrate similar performance in terms of estimate accuracy with similar log(MSE) values. From Table 3, the following conclusions could be drawn. First, for the *computation cost*, we find that (1) T_1 for the DMGD method is much larger than that for the other methods, because it requires a considerable number of iterations to converge; (2) the T_1 value of CSL increases dramatically as p increases due to Hessian inverse calculation with complexity of $O(p^3)$; and (3) the DQN methods perform well with the lowest T_1 values, which is especially true for large p. Second, for the *communication cost*, we find that the T_2 value for the DOSN method is the highest. This is as expected because DOSN needs to transfer a Hessian matrix for calculation. This leads to a complexity of order $O(p^2)$. In contrast, the DQN has the lowest T_2 value. Finally, in comparison of the *total time cost*, the DQN methods perform the best in terms of T. To summarise, the DQN methods demonstrate comparable estimation accuracy and the lowest total time cost.

3.3 Ultrahigh-dimensional features

We next consider the ultrahigh-dimensional feature situation with $p \gg n$. In this case, appropriate sparse structure has to be assumed to the true regression coefficient (Fan & Lv, 2008). Therefore, various screening methods (Fan & Lv, 2008; Fan & Song, 2010; He et al., 2013; G. Li et al., 2012; X. Li et al., 2020) can be readily applied but in a distributed way. Once the feature dimension is significantly reduced, the DQN algorithm can be readily applied. For the purpose of illustration, we consider here the sure independence screening method for generalised linear models (Fan & Song, 2010) and calculate the statistic in a distributed way as follows.

We start with a simulation setup as suggested by Fan and Song (2010). More specifically, the covariates are generated by $X_{ij} = (\varepsilon_{ij} + a_{ij}\varepsilon)(1 + a_{ij}^2)^{-1/2}$, where ε and $\{\varepsilon_{ij}\}_{j=1}^{[p/3]}$ are independently and identically distributed with $N(0, 1)$, $\{\varepsilon_{ij}\}_{j=[p/3]+1}^{[2p/3]}$ are independently and identically distributed following a double exponential distribution with location and scale parameters to be 0 and 1, and $\{\varepsilon_{ij}\}_{j=[2p/3]+1}^{p}$ are independently and identically distributed following a mixture normal distribution with equal weights on $N(-1, 1)$ and $N(1, 0.5)$. The $\{a_{ij}\}_{j=1}^{q}$ are independently and identically distributed with $N(0, 1)$ for the first q variables and $a_j = 0$ for $j \geq q$. The true feature set is $\mathcal{M}_T = \{1, \ldots, s\}$ with $s = 20$. Define $\theta^\top = (\theta_j) = \mathbf{1}_{[s/5]} \otimes (1, -1.1, 1.2, -1.3, 1.4)/2$, where $\mathbf{1}_b \in \mathbb{R}^b$ is a vector with all elements equal to 1, and $\otimes$ denotes the Kronecker product. The response Y_i is generated by a standard logistic regression. The feature dimension p and total sample size N are set at 10^4 and 10^5, respectively. The number of workers is set to $M = 20$, 40, and 50.

Next, we follow Fan and Song (2010) and compute the marginal maximum likelihood estimator for each feature j on the mth worker as $\widehat{\theta}_{j,m}$. In most cases, this should be a biased estimate for θ_j, but could be useful for variable screening. This leads to a total of M marginal estimators $\widehat{\theta}_{j,m}$. These are then averaged as $\widetilde{\theta}_j = M^{-1} \sum_m \widehat{\theta}_{j,m}$, which is an overall marginal estimator for θ_j. We

Table 4. Log(MSE) values and corresponding SD, IQR, and range for ultrahigh-dimensional case

			Stage 0		Stage 1		Stage 2		Stage 3		Stage 4		
	M	CR	SR1	BFGS	SR1	BFGS	SR1	BFGS	SR1	BFGS	SR1	BFGS	MLE
Log	2	1.00	0.34	0.34	−1.01	−0.95	−0.93	−0.86	−0.99	−1.01	−1.11	−1.08	−1.00
(MSE)	4	1.00	0.35	0.35	−1.34	−0.87	−1.29	−1.11	−1.42	−1.42	−1.47	−1.49	−1.45
	5	1.00	0.40	0.40	−1.37	−0.73	−1.38	−1.13	−1.56	−1.53	−1.61	−1.64	−1.60
SD	2	1.00	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.05
	4	1.00	0.06	0.06	0.06	0.05	0.07	0.07	0.06	0.06	0.06	0.06	0.06
	5	1.00	0.07	0.07	0.06	0.08	0.08	0.09	0.07	0.07	0.07	0.07	0.07
IQR	2	1.00	0.08	0.08	0.07	0.05	0.06	0.08	0.07	0.07	0.06	0.06	0.07
	4	1.00	0.09	0.09	0.09	0.08	0.10	0.10	0.10	0.10	0.09	0.09	0.09
	5	1.00	0.09	0.09	0.08	0.09	0.12	0.13	0.10	0.11	0.10	0.09	0.10
Range	2	1.00	0.25	0.25	0.24	0.19	0.24	0.25	0.22	0.22	0.24	0.24	0.25
	4	1.00	0.27	0.27	0.26	0.25	0.34	0.36	0.30	0.32	0.30	0.30	0.29
	5	1.00	0.30	0.30	0.30	0.37	0.38	0.46	0.34	0.37	0.34	0.33	0.32

Note. The numerical performance is evaluated for different M (×10) and methods. The whole sample size N and feature dimension p are fixed at 10^5 and 10^4, respectively. Finally, the reported results are averaged based on $R = 100$ simulations.
MSE = mean squared error; SD = standard deviation; IQR = inter-quartile range; CR = coverage rate; BFGS = Broyden–Fletcher–Goldfarb–Shanno.

next obtain the estimated feature set as $\widetilde{\mathcal{M}} = \{1 \le j \le p : |\widetilde{\theta}_j| \ge \gamma_n\}$, where γ_n is appropriately selected such that $|\widetilde{\mathcal{M}}| = \lceil n/\log(n) \rceil$; see Fan and Song (2010) for a more detailed discussion. Consequently, the condition (C6) for the DQN algorithm is automatically satisfied. Thereafter, the proposed DQN method can be readily applied to the dimension reduced problem with only the selected feature involved.

To measure the performance of the distributed screening procedure and the DQN algorithm, we compute the coverage rate for the rth $(1 \le r \le R)$ replication as $\mathrm{CR}^{(r)} = |\widetilde{\mathcal{M}}^{(r)} \bigcap \mathcal{M}_T^{(r)}|/|\mathcal{M}_T^{(r)}|$. Then, the overall coverage rate is given by $\mathrm{CR} = R^{-1} \sum_r \mathrm{CR}^{(r)}$. The other metrics used in Section 3.1 are also considered. The detailed results are given in Table 4. From Table 4, we find that the implemented screening procedure is screening consistent in the sense that all CR values are equal to 1. Furthermore, with a fixed N, we find that a larger M always leads to a smaller n. This leads to a smaller screening feature set with size $\lceil n/\log(n) \rceil$. Consequently, fewer redundant features are included. This further results in even smaller log(MSE) values. Lastly, for the DQN algorithms, a slightly larger number of stages (i.e. larger K) are required to obtain an estimator that is competitive with MLE. These results are consistent with our theoretical findings in Theorems 1 and 2 and Corollary 1.

3.4 Real data analysis

In this section, we apply the proposed method to the THU Chinese news dataset for illustration. The dataset is publicly available at http://thuctc.thunlp.org. The dataset consists of 14 types of Chinese news collected from Sina news (https://news.sina.com.cn) from 2005 to 2011.

For the purpose of illustration, we generate response Y_i as follows. We first select all the news of type *technology* and define the response $Y_i = 1$. This leads to a total of $N_p = 162,929$ positive cases. We next randomly sample a total of $\lceil 1.5N_p \rceil$ negative cases without replacement from the other types of news. The corresponding response Y_i is defined to be 0. Then, the total sample size is given by $N = 407,322$. Different words are then extracted from the original documents. Those words with top $\mathscr{F}\%$ frequencies for each class are selected. They are then coded as binary %covariates. We consider $\mathscr{F}\% = 0.3\%$, 0.4%, and 0.5%, which leads to $p = 998$, 1,333, and 1,660, respectively. All data are then randomly shuffled and distributed to $M = 20$ worker computers. The competing methods and performance measures remain the same as those in Section

Table 5. Averaged computation cost T_1, communication cost T_2, and total time cost T for the THU Chinese news dataset

	p	DOSN	CSL	DMGD	DAQN	DQN-BFGS	DQN-SR1
T_1	998	1.22	3.54	104.93	13.34	0.72	0.95
	1,333	2.14	5.44	126.40	20.59	1.29	1.66
	1,660	2.63	8.69	151.70	28.84	1.96	2.39
T_2	998	62.56	0.36	38.52	0.12	0.99	0.71
	1,333	129.28	0.45	54.27	0.26	1.26	0.92
	1,660	193.91	0.56	66.65	0.36	1.57	1.10
T	998	63.78	3.90	143.45	13.46	1.71	1.66
	1,333	131.42	5.88	180.67	20.84	2.55	2.58
	1,660	196.54	9.25	218.36	29.20	3.53	3.49

Note. The time cost is evaluated for different methods with different feature dimensions p. The whole sample size N and number of workers M are fixed to be 407, 322 and 20, respectively. The reported results are averaged for $R = 10$ simulation replications.
DOSN = distributed one-step Newton; CSL = communication-efficient surrogate likelihood; DMGD = distributed momentum gradient descent; DAQN = distributed asynchronous averaged quasi-Newton; DQN = distributed quasi-Newton; BFGS = Broyden–Fletcher–GoldfarbShanno.

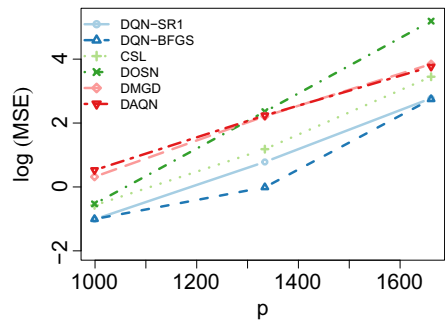

Figure 3. Log(MSE) values for the THU Chinese news dataset. The log(MSE) values are evaluated for different methods with different dimension p. The whole sample size N is fixed ta $N = 407, 322$ and number of workers M is fixed at $M = 20$. Finally, the reported log(MSE) values are averaged for $R = 10$ simulations. MSE = mean squared error.

3.2 but with $K = 5$. Because we do not know the ground truth in real data analysis, the global estimators are then treated as if they were the true parameters. The experiment is randomly replicated $R = 10$ times for a reliable evaluation. The results are summarised in Table 5 and Figure 3.

From Table 5, we find that the proposed DQN method has the lowest computation cost T_1. It outperforms other competing methods significantly in terms of computation efficiency. The computation advantage is particularly apparent when the feature dimension p is relatively large. Moreover, we find that the communication cost T_2 of the DQN methods is slightly higher than the smallest T_2 value for the DAQN method. However, the overall time cost of DQN (i.e. T) remains the smallest. This suggests that the proposed DQN methods are computationally very competitive. From Figure 3, we find that the proposed DQN methods also outperform their competitors slightly in terms of estimation accuracy with the smallest log(MSE) values.

4 Concluding remarks

This article focuses on the discussion of statistical properties of DQN algorithms, which is motivated by two well-known quasi-Newton algorithms, i.e. SR1 and BFGS. The proposed algorithms

are highly efficient both communicationally and computationally. We theoretically show that under mild conditions, only a small number of iterations are needed to obtain an estimator as statistically efficient as the global one. As far as we know, this is the first work to discuss the statistical properties of the DQN methods. Extensive numerical studies conducted on both simulation and real datasets are presented to illustrate the finite sample performance. To conclude this work, we discuss some interesting topics for future study. First, the DQN method proposed here requires that data among different worker computers are homogenous. This requirement may be difficult to satisfy for certain applications. Therefore, solving this problem should be an exciting topic for future research. In addition, the proposed algorithm ignores privacy in inter-computer communication. This could be of great concern when sensitive information needs to be transferred. Conducting DQN while ensuring data privacy will be investigated in the future.

Conflict of interest: None declared.

Funding

Shuyuan Wu's research is partially supported by the Shanghai Research Center for Data Science and Decision Technology. Danyang Huang's research is partially supported by the National Natural Science Foundation of China (No. 12071477), fund for building world-class universities (disciplines) of Renmin University of China, and Public Computing Cloud, Renmin University of China. Hansheng Wang's research is partially supported by the National Natural Science Foundation of China (No. 11831008) and the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (KLATASDS-MOE-ECNU-KLATASDS2101).

Data availability

The datasets were derived from sources in the public domain: the official website of THU Chinese Text Classification Package (http://thuctc.thunlp.org).

Appendices

Appendix A: Proof of the Main Theoretical Results

For simplicity, we define the following notation in the proof. Define $y_t = \dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},t+1}) - \dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},t})$ and $s_t = \widehat{\theta}_{\text{stage},t+1} - \widehat{\theta}_{\text{stage},t}$ for $t \geq 1$; particularly, $y_0 = \dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) - \dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})$ and $s_0 = \widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{stage},0}$. In addition, for the BFGS updating, according to (2.2), define $H_{t+1} = V_t^{\top} H_t V_t + \rho_t s_t s_t^{\top}$, where $V_t = I_p - \rho_t y_t s_t^{\top}$ and $\rho_t = 1/(s_t^{\top} y_t)$. For SR1 updating, according to (2.1), define $H_{t+1} = H_t + (v_t^{\top} y_t)^{-1}(v_t v_t^{\top})$, for $t \geq 1$, where $v_t = s_t - H_t y_t$. Particularly, $H_0 = M^{-1} \sum_{m=1}^{M} H_{(m,0)}$.

A.1 Proof of Theorem 1

We decompose the theorem proof into two parts. In the first part, we show that the distance between $\widehat{\theta}_{\text{stage},1}$ and $\widehat{\theta}_{\text{ge}}$ is bounded by $\kappa(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)} - \theta_0\|^2 + \|\ddot{\mathcal{L}}_{(m)}(\theta_0) - \Omega(\theta_0)\|_2^2 + \|\{\ddot{\mathcal{L}}_{(m)}(\theta_0) - \dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)} - \theta_0) \otimes I_p\}\|_2] + \|\widehat{\theta}_{\text{stage},0} - \theta_0\|)\|\widehat{\theta}_{\text{stage},0} - \theta_0\|$ with probability tending to 1. In the second part, we verify that when $N(p \log p)^2/n^4 \to 0$, then $\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\| = o_p(N^{-1/2})$.

Part 1. To analyse $\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|$, we first define the following 'good events':

$$\begin{aligned}
\mathcal{E}_0 &= \left\{ \|\widehat{\theta}_{\text{ge}} - \theta_0\| \leq \frac{\tau_{\min}}{4C_{\max}} \right\} \\
\mathcal{E}_m &= \left\{ \|\widehat{\theta}_{(m)} - \theta_0\| \leq \min\left\{ \frac{\tau_{\min}}{4C_{\max}}, \delta \right\}, n^{-1} \sum_{i \in \mathcal{S}_m} C(X_i, Y_i) \leq 2C_{\max}, \right. \\
&\qquad \left. \|\ddot{\mathcal{L}}_{(m)}(\theta_0) - \Omega(\theta_0)\|_2 \leq \frac{\delta\tau_{\min}}{4}, \|\dot{\mathcal{L}}_{(m)}(\theta_0)\| \leq \frac{(1-\delta)\tau_{\min}\delta_{\min}}{4} \right\}
\end{aligned} \tag{A.1}$$

where $\delta_{\min} = \min\{\delta, \delta\tau_{\min}/(4C_{\max})\}$. By Lemma 2, we know $P(\bigcup_{m=0}^{M} \mathcal{E}_m^c) \to 0$. In addition, it could be verified that the events $\bigcap_{m=0}^{M} \mathcal{E}'_m$ defined in Lemma 3 hold under $\bigcap_{m=0}^{M} \mathcal{E}_m$. Thus, it suffices to analyse the upper bound of $\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|$ under $\bigcap_{m=0}^{M} \mathcal{E}_m$ and $\bigcap_{m=0}^{M} \mathcal{E}'_m$.

We then proceed to study Part 1. Recall the definition of $\widehat{\theta}_{\text{stage},1}$, then by (C5), we have $\widehat{\theta}_{\text{stage},1} = \widehat{\theta}_{\text{stage},0} - M^{-1}\sum_{m=1}^{M}\{\ddot{\mathcal{L}}_{(m)}(\widehat{\theta}_{(m)})\}^{-1}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})$. In addition, define $\widehat{\theta}_{\text{nr},1} = \widehat{\theta}_{\text{stage},0} - \{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})\}^{-1}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})$ to represent the one-step Newton–Raphson estimator. Then, by the triangle inequality, we have

$$\begin{aligned}
\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\| &\le \|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{nr},1}\| + \|\widehat{\theta}_{\text{nr},1} - \widehat{\theta}_{\text{ge}}\| \\
&= \Bigg\| \widehat{\theta}_{\text{stage},0} - M^{-1}\sum_{m=1}^{M}\{\ddot{\mathcal{L}}_{(m)}(\widehat{\theta}_{(m)})\}^{-1}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0}) \\
&\quad - \Big[\widehat{\theta}_{\text{stage},0} - \{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})\}^{-1}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})\Big]\Bigg\| + \|\widehat{\theta}_{\text{nr},1} - \widehat{\theta}_{\text{stage},1}\| \\
&= \left\| \left[\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})\}^{-1} - M^{-1}\sum_{m=1}^{M}\{\ddot{\mathcal{L}}_{(m)}(\widehat{\theta}_{(m)})\}^{-1} \right] \dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0}) \right\| + \|\widehat{\theta}_{\text{nr},1} - \widehat{\theta}_{\text{ge}}\|
\end{aligned}$$

Denote $\Delta_1 = \{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})\}^{-1} - M^{-1}\sum_{m=1}^{M}\{\ddot{\mathcal{L}}_{(m)}(\widehat{\theta}_{(m)})\}^{-1}$. We then investigate Δ_1, $\widehat{\theta}_{\text{nr},1} - \widehat{\theta}_{\text{ge}}$, and $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})$ in the following three steps, respectively.

Step 1. By the triangle inequality, we have $\|\Delta_1\|_2 \le \|\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})\}^{-1} - \{\ddot{\mathcal{L}}(\theta_0)\}^{-1}\|_2 + \|\{\ddot{\mathcal{L}}(\theta_0)\}^{-1} - M^{-1}\sum_{m=1}^{M}\{\ddot{\mathcal{L}}_{(m)}(\theta_0)\}^{-1}\|_2 + \|M^{-1}\sum_{m=1}^{M}\{\ddot{\mathcal{L}}_{(m)}(\theta_0)\}^{-1} - M^{-1}\sum_{m=1}^{M} \{\ddot{\mathcal{L}}_{(m)}(\widehat{\theta}_{(m)})\}^{-1}\|_2 := \|\Delta_1^{(1)}\|_2 + \|\Delta_1^{(2)}\|_2 + \|\Delta_1^{(3)}\|_2$. We proceed to calculate the three terms separately.

Step 1.1. First, for any matrix B, we have $\|(B+\Delta B)^{-1} - B^{-1}\|_2 \le \|B^{-1}\|_2^2\|\Delta B\|_2$ (Jordan et al., 2019). Substituting $B = \ddot{\mathcal{L}}(\theta_0)$ and $\Delta B = \ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0}) - \ddot{\mathcal{L}}(\theta_0)$, it could be shown that

$$\|\Delta_1^{(1)}\|_2 \le \left\|\{\ddot{\mathcal{L}}(\theta_0)\}^{-1}\right\|_2^2 \left\|\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0}) - \ddot{\mathcal{L}}(\theta_0)\right\|_2 \le \frac{4}{(1-\delta)^2\tau_{\min}^2} 2C_{\max}\|\widehat{\theta}_{\text{stage},0} - \theta_0\|$$

The second inequality holds because $\|\{\ddot{\mathcal{L}}(\theta_0)\}^{-1}\|_2 \le 2/\{(1-\delta)\tau_{\min}\}$ under $\bigcap_{m=0}^{M}\mathcal{E}'_m$ and $\|\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0}) - \ddot{\mathcal{L}}(\theta_0)\|_2 \le 2C_{\max}\|\widehat{\theta}_{\text{stage},0} - \theta_0\|$ under $\bigcap_{m=0}^{M}\mathcal{E}_m$. Consequently, there exists a constant $\kappa > 0$ such that $\|\Delta_1^{(1)}\|_2 \le \kappa\|\widehat{\theta}_{\text{stage},0} - \theta_0\|/(6\times 2C_{\max})$.

Step 1.2 Next, we analyse $\Delta_1^{(2)}$. It could be shown that

$$\begin{aligned}
\Delta_1^{(2)} &= M^{-1}\sum_{m=1}^{M}\{\ddot{\mathcal{L}}_{(m)}(\theta_0)\}^{-1}\{\ddot{\mathcal{L}}(\theta_0) - \ddot{\mathcal{L}}_{(m)}(\theta_0)\}\{\ddot{\mathcal{L}}(\theta_0)\}^{-1} \\
&= M^{-1}\sum_{m=1}^{M}\Big(\big[\{\ddot{\mathcal{L}}_{(m)}(\theta_0)\}^{-1} - \{\ddot{\mathcal{L}}(\theta_0)\}^{-1}\big]\{\ddot{\mathcal{L}}(\theta_0) - \ddot{\mathcal{L}}_{(m)}(\theta_0)\}\{\ddot{\mathcal{L}}(\theta_0)\}^{-1} \\
&\quad + \{\ddot{\mathcal{L}}(\theta_0)\}^{-1}\{\ddot{\mathcal{L}}(\theta_0) - \ddot{\mathcal{L}}_{(m)}(\theta_0)\}\{\ddot{\mathcal{L}}(\theta_0)\}^{-1}\Big) \\
&= M^{-1}\sum_{m=1}^{M}\{\ddot{\mathcal{L}}_{(m)}(\theta_0)\}^{-1}\{\ddot{\mathcal{L}}(\theta_0) - \ddot{\mathcal{L}}_{(m)}(\theta_0)\}\{\ddot{\mathcal{L}}(\theta_0)\}^{-1}\{\ddot{\mathcal{L}}(\theta_0) - \ddot{\mathcal{L}}_{(m)}(\theta_0)\}\{\ddot{\mathcal{L}}(\theta_0)\}^{-1}
\end{aligned}$$

We then have $\|\Delta_1^{(2)}\|_2 \le \|\{\ddot{\mathcal{L}}_{(m)}(\theta_0)\}^{-1}\|_2\|\{\ddot{\mathcal{L}}(\theta_0)\}^{-1}\|_2^2 \times M^{-1}\sum_{m=1}^{M}\|\ddot{\mathcal{L}}(\theta_0) - \ddot{\mathcal{L}}_{(m)}(\theta_0)\|_2^2 \le 6\{(1-\delta)^3 \tau_{\min}^3\}^{-1} \times 8M\sum_{m=1}^{M}\|\ddot{\mathcal{L}}_{(m)}(\theta_0) - \Omega(\theta_0)\|_2^2$. This is because $\|\ddot{\mathcal{L}}(\theta_0) - \ddot{\mathcal{L}}_{(m)}(\theta_0)\|_2^2 \le$

$2\{\|\ddot{\mathcal{L}}(\theta_0)-\Omega(\theta_0)\|_2^2+\|\ddot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2\}$ and $\|\ddot{\mathcal{L}}(\theta_0)-\Omega(\theta_0)\|_2^2 \le \|M^{-1}\sum_{m=1}^{M}(\ddot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0))\|_2^2 \le (1+1/M)M^{-1}\sum_{m=1}^{M}\|\ddot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2$. Therefore, there exists a constant $\kappa>0$ such that

$$\left\|\Delta_1^{(2)}\right\|_2 \le \frac{\kappa}{6\times 2C_{\max}} M^{-1}\sum_{m=1}^{M}\|\ddot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2$$

Step 1.3. Moreover, it could be proved that

$$\Delta_1^{(3)} = M^{-1}\sum_{m=1}^{M}\left\{\ddot{\mathcal{L}}_{(m)}(\widehat{\theta}_{(m)})\right\}^{-1}\left\{\ddot{\mathcal{L}}_{(m)}(\widehat{\theta}_{(m)})-\ddot{\mathcal{L}}_{(m)}(\theta_0)\right\}\left\{\ddot{\mathcal{L}}_{(m)}(\theta_0)\right\}^{-1}$$

By Taylor's expansion, Cauchy–Schwarz inequality, and *Step 1.2*, we have

$$\begin{aligned}\|\Delta_1^{(3)}\|_2 \le{}& \left\| M^{-1}\sum_{m=1}^{M}\left\{\ddot{\mathcal{L}}(\theta_0)\right\}^{-1}\left\{\ddot{\mathcal{L}}_{(m)}(\widehat{\theta}_{(m)})-\ddot{\mathcal{L}}_{(m)}(\theta_0)\right\}\left\{\ddot{\mathcal{L}}(\theta_0)\right\}^{-1}\right\|_2 \\ &+\frac{\kappa}{6\times 2C_{\max}} M^{-1}\sum_{m=1}^{M}\left\{\|\ddot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2+\|\widehat{\theta}_{(m)}-\theta_0\|^2\right\}\end{aligned} \tag{A.2}$$

Hence, it suffices to study the first term of (A.2). Using Taylor's expansion again, it could be verified that

$$\begin{aligned}\ddot{\mathcal{L}}_{(m)}(\widehat{\theta}_{(m)})-\ddot{\mathcal{L}}_{(m)}(\theta_0) &= \dddot{\mathcal{L}}_{(m)}(\theta_0)\left\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\right\}+\left\{\dddot{\mathcal{L}}_{(m)}(\xi_{(m)})-\dddot{\mathcal{L}}_{(m)}(\theta_0)\right\}\left\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\right\} \\ &= \left\{\dddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\right\}\left\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\right\}+\dot{\Omega}(\theta_0)\left\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\right\}+\mathcal{O}\end{aligned}$$

where $\xi_{(m)}=\eta_{(m)}\widehat{\theta}_{(m)}+(1-\eta_{(m)})\theta_0$ for some $0\le\eta_{(m)}\le 1$, $\mathcal{O}=\{\dddot{\mathcal{L}}_{(m)}(\xi_{(m)})-\dddot{\mathcal{L}}_{(m)}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}$. In addition, we have $\|\mathcal{O}\|_2 \le 2C_{\max}\|\widehat{\theta}_{(m)}-\theta_0\|^2$ by (C4). Replacing the results back into (A.2), we obtain $\|\Delta_1^{(3)}\|_2 \le \kappa(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)}-\theta_0\|^2+\|\ddot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2+\|\{\dddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}\|_2]+\|\widehat{\theta}_{\text{stage},0}-\theta_0\|)/(6\times 2C_{\max})$. Combining the above-mentioned results, we have $\|\Delta_1\|_2 \le \kappa(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)}-\theta_0\|^2+\|\ddot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2+\|\{\dddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}\|_2]+\|\widehat{\theta}_{\text{stage},0}-\theta_0\|)/(4C_{\max})$. This finishes the proof of *Step 1*.

Step 2. In this step, we study $\widehat{\theta}_{\text{nr},1}-\widehat{\theta}_{\text{ge}}$ and $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})$. From Theorem 5.3 in Bubeck (2015), when $\|\widehat{\theta}_{\text{stage},0}-\widehat{\theta}_{\text{ge}}\| \le \lambda_{\min}\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}/(2C_{\text{ge}})$, where C_{ge} is the global Lipschitz constant of $\ddot{\mathcal{L}}(\theta)$ such that $\|\ddot{\mathcal{L}}(\theta')-\ddot{\mathcal{L}}(\theta'')\| \le C_{\text{ge}}\|\theta'-\theta''\|$, we have

$$\|\widehat{\theta}_{\text{nr},1}-\widehat{\theta}_{\text{ge}}\| \le \frac{C_{\text{ge}}}{\lambda_{\min}\left\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\right\}}\|\widehat{\theta}_{\text{stage},0}-\widehat{\theta}_{\text{ge}}\|^2 \le \frac{4C_{\max}}{(1-\delta)\tau_{\min}}\|\widehat{\theta}_{\text{stage},0}-\widehat{\theta}_{\text{ge}}\|^2 \tag{A.3}$$

Moreover, by (C4), it could be verified that $\|\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})-\dot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\| \le 2C_{\max}\|\widehat{\theta}_{\text{stage},0}-\widehat{\theta}_{\text{ge}}\|$. This finishes the proof of *Step 2*.

Combining the results of *Steps 1* and *2*, we have $\|\widehat{\theta}_{\text{stage},1}-\widehat{\theta}_{\text{ge}}\| \le \kappa(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)}-\theta_0\|^2+\|\ddot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2+\|\{\dddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}\|_2]+\|\widehat{\theta}_{\text{stage},0}-\theta_0\|+\|\widehat{\theta}_{\text{ge}}-\theta_0\|)\ \|\widehat{\theta}_{\text{stage},0}-\widehat{\theta}_{\text{ge}}\|$ with probability tending to 1. Noting that $\|\widehat{\theta}_{\text{ge}}-\theta_0\|$ is a negligible higher-order term, and we finish the first part.

Part 2. To prove the second part, we separately analyse the convergence properties of $M^{-1}\sum_{m=1}^{M}\|\widehat{\theta}_{(m)}-\theta_0\|^2$, $M^{-1}\sum_{m=1}^{M}\|\ddot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2$, $M^{-1}\sum_{m=1}^{M}\{\dddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}$ and $\|\widehat{\theta}_{\text{stage},0}-\theta_0\|^2$. By Lemma 1 we know

$$E\|\widehat{\theta}_{(m)}-\theta_0\|^2 \le C_1 n^{-1} C_G^2\{1+o(1)\}$$

$$E\left\{\|\ddot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2\right\} \le C_2\frac{\log p}{n}\{1+o(1)\}$$

$$E\left\|\{\ddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}\right\|_2 \le C_4\frac{p\sqrt{\log p}}{n}\{1+o(1)\}$$

$$E\left[\|\widehat{\theta}_{\text{stage},0}-\theta_0\|^2\right] \le \left(\frac{2C_G^2}{\tau_{\min}^2 N}+\frac{C_3 C_G^2 C_H^2\log p}{\tau_{\min}^4 n^2}\right)\{1+o(1)\}$$

for some positive constants C_1–C_4. Moreover, by Markov's inequality, we have

$$M^{-1}\sum_{m=1}^{M}\left\{\|\widehat{\theta}_{(m)}-\theta_0\|^2+\|\ddot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2\right\}+\left\|\{\ddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}\right\|_2$$
$$+\|\widehat{\theta}_{\text{stage},0}-\theta_0\| = O_p(n^{-1}p\sqrt{\log p}+N^{-1/2}) \quad \text{and} \quad \|\widehat{\theta}_{\text{stage},0}-\theta_0\| = O_p(1/\sqrt{N}+\sqrt{\log p}/n) \tag{A.4}$$

Hence, $\|\widehat{\theta}_{\text{stage},1}-\widehat{\theta}_{\text{ge}}\| = O_p(n^{-2}p\log p)+o_p(N^{-1/2})$. Furthermore, under the condition $N(p\log p)^2/n^4\to 0$, we have $N^{1/2}\|\widehat{\theta}_{\text{stage},1}-\widehat{\theta}_{\text{ge}}\|\to_p 0$, which finishes the proof of the second part, thereby completing the proof of the entire theorem.

A.2 Proof of Theorem 2

To verify Theorem 2, we first prove that $\|\widehat{\theta}_{\text{stage},2}-\widehat{\theta}_{\text{ge}}\| \le \kappa_2(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)}-\theta_0\|^2+\|\ddot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2+\|\{\ddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}\|_2]+\|\widehat{\theta}_{\text{stage},0}-\theta_0\|)\|\widehat{\theta}_{\text{stage},1}-\widehat{\theta}_{\text{ge}}\|$ for some constant $\kappa_2>0$, with probability tending to 1. Next, we verify the optimality of $\widehat{\theta}_{\text{stage},2}$ under the condition $Np^4(\log p)^3/n^6\to 0$.

Note that by Algorithms 2 and 4, the proposed methods realise the global update of the approximated Hessian inverse. In other words, the two-stage estimator update (2.4) is equal to $\widehat{\theta}_{\text{stage},2}=\widehat{\theta}_{\text{stage},1}-H_1\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})$. For convenience, instead of directly studying $\widehat{\theta}_{\text{stage},1}-H_1\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})$, we investigate $\widehat{\theta}_{\text{stage},1}-(B_1)^{-1}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})$, where $B_1=H_1^{-1}$. By the triangle inequality, it could be verified that

$$\|\widehat{\theta}_{\text{stage},2}-\widehat{\theta}_{\text{ge}}\| \le \|\widehat{\theta}_{\text{stage},1}-\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})-\widehat{\theta}_{\text{ge}}\|+\left\|\left[\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1}-B_1^{-1}\right]\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})\right\|$$

We denote $\Delta_2^{(1)}=\widehat{\theta}_{\text{stage},1}-\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})-\widehat{\theta}_{\text{ge}}$ and $\Delta_2^{(2)}=[\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1}-B_1^{-1}]\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})$, where $\Delta_2^{(1)}$ is independent of SR1 or BFGS update. Furthermore, by similar analytical techniques as those used in (A.3), we have $\|\Delta_2^{(1)}\|\le\kappa_2'\|\widehat{\theta}_{\text{stage},1}-\widehat{\theta}_{\text{ge}}\|^2$ for some constant $\kappa_2'>0$ with probability tending to 1. Hence, it suffices to study $\Delta_2^{(2)}$. Therefore, we investigate $\Delta_2^{(2)}$ under the good events $\bigcap_{m=0}^{M}\mathcal{E}_m$ and $\mathcal{E}'$ by SR1 and BFGS update separately.

Part 1 (SR1). By the Sherman–Morrison formula (Burden et al., 2015, Theorem 10.8), the SR1 updating formula can be expressed as

$$B_1=B_0+\frac{(y_0-B_0s_0)(y_0-B_0s_0)^{\top}}{(y_0-B_0s_0)^{\top}y_0}$$

where $B_0=H_0^{-1}$. Then, we proceed to study $\Delta_2^{(2)}$, which can be rewritten as

$$\Delta_2^{(2)}=\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1}\{B_1-\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}B_1^{-1}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) \tag{A.5}$$

Furthermore, when $s_1 = B_1^{-1}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})$, by the triangle inequality, we have

$$\|\Delta_2^{(2)}\| \le \|\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1}\|_2\{\|y_1 - B_1 s_1\| + \|y_1 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})s_1\|\}$$

We then study $y_1 - B_1 s_1$ and $y_1 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})s_1$.

Step 1. First, we investigate $y_1 - B_1 s_1$. By (2.1), we have

$$y_1 - B_1 s_1 = y_1 - B_0 s_1 + \frac{r_0 r_0^\top s_1}{r_0^\top s_0}$$

where $r_t = y_t - B_t s_t$ for any $t > 0$. Then by Taylor's expansion, it could be proved that

$$\begin{aligned}\|y_1 - B_0 s_1\| &\le \Big\|(\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) - B_0)(\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{stage},1}) + \ddot{\mathcal{L}}(\xi_1)(\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{stage},1}) \\ &\quad - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})(\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{stage},1})\Big\| \\ &\le 2\Big\{\|\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) - B_0\|_2\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\| + \|\ddot{\mathcal{L}}(\xi_1) - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})\|_2\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|\Big\} \\ &\le 2\Big\{\|\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) - B_0\|_2 + 2C_{\max}\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|\Big\}\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|\end{aligned} \tag{A.6}$$

Here $\xi_1 = \eta_1\widehat{\theta}_{\text{stage},2} + (1-\eta_1)\widehat{\theta}_{\text{stage},1}$ with some $0 \le \eta_1 \le 1$. The second inequality in (A.6) holds by the triangle inequality $\|\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{stage},1}\| \le \|\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{ge}}\| + \|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|$ and Lemma 4. The last inequality in (A.6) holds by (C4). In addition, from (D.5), it could be verified that

$$\begin{aligned}|(r_0^\top s_0)^{-1}(r_0 r_0^\top s_1)| &\le \frac{\|r_0\|\|s_1\|}{c_1\|s_0\|} = \frac{\|\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) - \dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0}) - B_0(\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{stage},0})\|\|\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{stage},1}\|}{c_1\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{stage},0}\|} \\ &= \frac{\|\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) - B_0\}s_0 + \{\ddot{\mathcal{L}}(\xi_0) - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})\}s_0\|\|\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{stage},1}\|}{c_1\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{stage},0}\|} \\ &\le 2\Big\{\|\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) - B_0\|_2\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\| + 4C_{\max}\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|\|\widehat{\theta}_{\text{stage},0} - \widehat{\theta}_{\text{ge}}\|\Big\}\end{aligned} \tag{A.7}$$

Here $\xi_0 = \eta_0\widehat{\theta}_{\text{stage},1} + (1-\eta_1)\widehat{\theta}_{\text{stage},0}$ with some $0 \le \eta_0 \le 1$, and the last inequality holds by Lemma 4 and (C4). Combining the results of (A.6) and (A.7), we have $\|y_1 - B_1 s_1\| \le 4\{\|\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) - B_0\|_2 + 4C_{\max}\|\widehat{\theta}_{\text{stage},0} - \widehat{\theta}_{\text{ge}}\|\}\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|$. Furthermore, $\|\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) - B_0\|_2 \le \|\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})\|_2^2\|\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1})\}^{-1} - M^{-1}\sum_{m=1}^M \ddot{\mathcal{L}}_{(m)}(\widehat{\theta}_{(m)})\|_2$. Using an analysis technique similar to the one used in Appendix A.1 *Step 1* to study the value of Δ_1, we have $\|\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) - B_0\|_2 \le \kappa_2'(M^{-1}\sum_{m=1}^M[\|\widehat{\theta}_{(m)} - \theta_0\|^2 + \|\ddot{\mathcal{L}}_{(m)}(\theta_0) - \Omega(\theta_0)\|_2^2 + \|\{\ddot{\mathcal{L}}_{(m)}(\theta_0) - \dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)} - \theta_0) \otimes I_p\}\|_2] + \|\widehat{\theta}_{\text{stage},0} - \theta_0\|)$. Hence, we have $\|y_1 - B_1 s_1\| \le \kappa_2'(M^{-1}\sum_{m=1}^M[\|\widehat{\theta}_{(m)} - \theta_0\|^2 + \|\ddot{\mathcal{L}}_{(m)}(\theta_0) - \Omega(\theta_0)\|_2^2 + \|\{\ddot{\mathcal{L}}_{(m)}(\theta_0) - \dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)} - \theta_0) \otimes I_p\}\|_2] + \|\widehat{\theta}_{\text{stage},0} - \theta_0\| + \|\widehat{\theta}_{\text{ge}} - \theta_0\|)\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|$. This accomplishes the proof of *Step 1*.

Step 2. In this step, we show that $y_1 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})s_1$ is a negligible higher-order term. By Taylor's expansion, it could be proved that

$$\begin{aligned}\|y_1 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})s_1\| &= \|\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},2}) - \dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})(\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{stage},1})\| \\ &\le \|\{\ddot{\mathcal{L}}(\xi_1) - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}\|\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{stage},1}\| \le 4C_{\max}\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|^2\end{aligned} \tag{A.8}$$

The last inequality holds by Lemma 4 and (C4). Combining the results of Steps 1 and 2, we finish the first part of the theorem proof.

Part 2 (BFGS). Recall $\Delta_2^{(2)} = \{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1}\{B_1 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}B_1^{-1}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) = \{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1}\ \{B_1 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}s_1$. Denote

$$P_0 = I_p - \frac{\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{1/2}s_0\left[\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}y_0\right]^{\top}}{y_0^{\top}s_0}$$

Then it could be shown by Broyden et al. (1973, Lemma 5.1) that

$$E_1 = P_0^{\top}E_0P_0 + \frac{\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}\{y_0 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})s_0\}\left[\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}y_0\right]^{\top}}{y_0^{\top}s_0}$$
$$+ \frac{\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}y_0\{y_0 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})s_0\}^{\top}\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}P_0}{y_0^{\top}s_0}$$

where $E_0 = \{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}\{B_0 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}$ and $E_1 = \{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}\{B_1 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}$. Then it could be proved that

$$\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{1/2}\Delta_2^{(2)}$$
$$= E_1\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{1/2}s_1 = P_0^{\top}E_0P_0\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{1/2}s_1 + \frac{\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}\{y_0 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})s_0\}y_0^{\top}s_1}{y_0^{\top}s_0}$$
$$+ \frac{\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}y_0\{y_0 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})s_0\}^{\top}\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}P_0s_1}{y_0^{\top}s_0} := \Delta_2^{(2,1)} + \Delta_2^{(2,2)} + \Delta_2^{(2,3)}$$

Note that $\|y_0 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})s_0\| \le \|\widehat{\theta}_{\text{stage},0} - \widehat{\theta}_{\text{ge}}\|^2$ by (A.8). Applying similar analytical techniques as in Part 1, and using $|y_0s_0| = |s_0^{\top}\ddot{\mathcal{L}}(\xi_0)s_0| \ge c_1\|s_0\|^2$ for some positive constant $c_1 > 0$, it could be verified that

$$\|\Delta_2^{(2,2)}\| \le \frac{\|\widehat{\theta}_{\text{stage},0} - \widehat{\theta}_{\text{ge}}\|^2\|y_0\|\|s_1\|}{c_1\|s_0\|^2} \le \frac{\|\widehat{\theta}_{\text{stage},0} - \widehat{\theta}_{\text{ge}}\|^2\|\ddot{\mathcal{L}}(\xi_0)\|_2\|s_1\|}{c_1\left|\|\widehat{\theta}_{\text{stage},0} - \widehat{\theta}_{\text{ge}}\| - \|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|\right|}$$
$$\le \kappa_2'\|\widehat{\theta}_{\text{stage},0} - \widehat{\theta}_{\text{ge}}\|\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|$$

Similarly, for $\|P_0\| \le 1 + 1/c_1$, we have $\|\Delta_2^{(2,3)}\| \le \kappa_2'\|\widehat{\theta}_{\text{stage},0} - \widehat{\theta}_{\text{ge}}\|\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|$. Thus, it can be verified that $\|\Delta_2^{(2,1)}\| \le \kappa_2'\|B_0 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\|_2\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|$. By similar analysis of $\|\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0}) - B_0\|_2$ as in Appendix A.1 *Step 1*, we have $\|B_0 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\| \le \kappa_2'(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)} - \theta_0\|^2 + \|\ddot{\mathcal{L}}_{(m)}(\theta_0) - \Omega(\theta_0)\|_2^2 + \|\{\ddot{\mathcal{L}}_{(m)}(\theta_0) - \dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)} - \theta_0) \otimes I_p\}\|_2] + \|\widehat{\theta}_{\text{stage},0} - \theta_0\| + \|\widehat{\theta}_{\text{ge}} - \theta_0\|)$. Thus, $\Delta_2^{(2)}$ could be bounded by $\|\Delta_2^{(2)}\| \le \kappa_2'(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)} - \theta_0\|^2 + \|\ddot{\mathcal{L}}_{(m)}(\theta_0) - \Omega(\theta_0)\|_2^2 + \|\{\ddot{\mathcal{L}}_{(m)}(\theta_0) - \dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)} - \theta_0) \otimes I_p\}\|_2] + \|\widehat{\theta}_{\text{stage},0} - \theta_0\| + \|\widehat{\theta}_{\text{ge}} - \theta_0\|)\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|$. This finishes the proof of Part 2 (BFGS).

Next, by (A.4) again, we have $\|\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{ge}}\| = O_p(p^2(\log p)^{3/2}/n^3) + o_p(1/\sqrt{N})$. As a result, under the condition $N(p^4(\log p)^3)/n^6 \to 0$, we have $\|\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{ge}}\| = o_p(N^{-1/2})$, which accomplishes the whole theorem proof.

A.3 Proof of Corollary 1

First, to prove $\|\widehat{\theta}_{\text{stage},K} - \widehat{\theta}_{\text{ge}}\| \le \kappa_K(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)} - \theta_0\|^2 + \|\ddot{\mathcal{L}}_{(m)}(\theta_0) - \Omega(\theta_0)\|_2^2 + \|\{\ddot{\mathcal{L}}_{(m)}(\theta_0) - \dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)} - \theta_0) \otimes I_p\}\|_2] + \|\widehat{\theta}_{\text{stage},0} - \theta_0\|)^K\|\widehat{\theta}_{\text{stage},0} - \theta_0\|$, we verify that $\|\widehat{\theta}_{\text{stage},K} - \widehat{\theta}_{\text{ge}}\| \le$

$\kappa_K(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)}-\theta_0\|^2+\|\dot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2+\|\{\ddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}\|_2]+\|\widehat{\theta}_{\text{stage},0}-\theta_0\|)\|\widehat{\theta}_{\text{stage},K-1}-\widehat{\theta}_{\text{ge}}\|$. In addition, by the triangle inequality, we have

$$\begin{aligned}\|\widehat{\theta}_{\text{stage},k}-\widehat{\theta}_{\text{ge}}\| &= \|\widehat{\theta}_{\text{stage},k-1}-B_{k-1}^{-1}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},k-1})-\widehat{\theta}_{\text{ge}}\| \\ &\le \|\widehat{\theta}_{\text{stage},k-1}-\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},k-1})-\widehat{\theta}_{\text{ge}}\|+\|[\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1}-B_{k-1}^{-1}]\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},k-1})\| \end{aligned} \tag{A.9}$$

for any $2\le k\le K$. Similar to the analysis of (A.3) at the beginning of Appendix A.2, it could be proved that $\|\widehat{\theta}_{\text{stage},k-1}-\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},k-1})-\widehat{\theta}_{\text{ge}}\|\le\kappa'_{k-1}\|\widehat{\theta}_{\text{stage},k-1}-\widehat{\theta}_{\text{ge}}\|^2$ with probability tending to 1. Because the first term in (A.9) is a negligible higher-order term, it suffices to study the upper bound of the second term in (A.9).

Denote $\Delta_3=[\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1}-B_{k-1}^{-1}]\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},k-1})$; then it could be verified that

$$\begin{aligned}\|\Delta_3\| &= \|\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1}\{B_{k-1}^{-1}-\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}B_{k-1}^{-1}\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},k-1})\| \\ &\le \|\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1}\|_2\|B_{k-1}^{-1}-\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\|_2\|B_{k-1}^{-1}\|_2\|\ddot{\mathcal{L}}(\xi_{k-1})\|_2\|\widehat{\theta}_{\text{stage},k-1}-\widehat{\theta}_{\text{ge}}\|\end{aligned}$$

By Lemmas 3 and 4, we know $\|\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1}\|_2$ and $\|\ddot{\mathcal{L}}(\xi_{k-1})\|_2$ are both bounded by some constant $C>0$ with probability tending to 1. As a consequence, to prove Corollary 1, it suffices to prove that for any $2\le k\le K$, $\|B_k-\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\|\le\kappa_k(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)}-\theta_0\|^2+\|\dot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2+\|\{\ddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}\|_2]+\|\widehat{\theta}_{\text{stage},0}-\theta_0\|)$ with probability tending to 1. We then verify the inequality by the inductive method under the SR1 and BFGS update separately as follows.

Part 1 (SR1). Assume that $\|B_{k-1}-\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\|_2\le\kappa_{k-1}(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)}-\theta_0\|^2+\|\dot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2+\|\{\ddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}\|_2]+\|\widehat{\theta}_{\text{stage},0}-\theta_0\|)$. The goal is to verify that $\|B_k-\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\|_2\le\kappa_k(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)}-\theta_0\|^2+\|\dot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2+\|\{\ddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}\|_2]+\|\widehat{\theta}_{\text{stage},0}-\theta_0\|)$. To this end, by the SR1 updating formula and (D.5), we have

$$\|B_k-\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\|_2\le\|B_{k-1}-\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\|_2+\frac{\|r_{k-1}\|}{c_1\|s_{k-1}\|}$$

Furthermore, it could be proved that, with probability tending to 1, we have

$$\begin{aligned}\frac{\|r_{k-1}\|}{c_1\|s_{k-1}\|} &\le \frac{\|\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},k-1})-B_{k-1}\|_2\|s_{k-1}\|+\|\ddot{\mathcal{L}}(\xi_{k-1})-\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},k-1})\|_2\|s_{k-1}\|}{c_1\|\widehat{\theta}_{\text{stage},k}-\widehat{\theta}_{\text{stage},k-1}\|} \\ &\le c_1^{-1}\left\{\|\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})-B_{k-1}\|_2+\|\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},k-1})-\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\|_2+\|\ddot{\mathcal{L}}(\xi_{k-1})-\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},k-1})\|_2\right\} \\ &\le \kappa_k\Bigg(M^{-1}\sum_{m=1}^{M}\Big[\|\widehat{\theta}_{(m)}-\theta_0\|^2+\|\dot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2+\|\widehat{\theta}_{\text{stage},0}-\theta_0\| \\ &\quad+\{\ddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}\Big]\Bigg)\end{aligned}$$

The last inequality holds because $\|\widehat{\theta}_{\text{stage},k-1}-\widehat{\theta}_{\text{ge}}\|$ is a higher-order term compared to $M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)}-\theta_0\|^2+\|\dot{\mathcal{L}}_{(m)}(\theta_0)-\Omega(\theta_0)\|_2^2+\|\{\ddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}\|_2]+\|\widehat{\theta}_{\text{stage},0}-\theta_0\|$. This finishes the proof of the first part.

Part 2 (BFGS). Similar to the proof of Part 1, assume that $\|B_{k-1} - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\|_2 \leq \kappa_{k-1}(M^{-1}\sum_{m=1}^{M}[\|\widehat{\theta}_{(m)} - \theta_0\|^2 + \|\ddot{\mathcal{L}}_{(m)}(\theta_0) - \Omega(\theta_0)\|_2^2 + \|\{\ddot{\mathcal{L}}_{(m)}(\theta_0) - \dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)} - \theta_0) \otimes I_p\}\|_2] + \|\widehat{\theta}_{\text{stage},0} - \theta_0\|)$. By the BFGS updating formula and Broyden et al. (1973, Lemma 5.1), we have

$$E_k = P_{k-1}^{\top} E_{k-1} P_{k-1} + \frac{\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}\{y_{k-1} - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})s_{k-1}\}\left[\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})^{-1/2} y_{k-1}\right]^{\top}}{y_{k-1}^{\top} s_{k-1}}$$
$$+ \frac{\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2} y_{k-1}\{y_{k-1} - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})s_{k-1}\}^{\top}\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2} P_{k-1}}{y_{k-1}^{\top} s_{k-1}}$$

where

$$P_{k-1} = I_p - \frac{\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{1/2} s_{k-1}\left[\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2} y_{k-1}\right]^{\top}}{y_{k-1}^{\top} s_{k-1}}$$

$E_k = \{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}\{B_k - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}$, and $E_{k-1} = \{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}\{B_{k-1} - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}^{-1/2}$. Then, using similar analytical techniques as in Part 2 of Appendix A.2, and that $|y_{k-1}^{\top} s_{k-1}| \geq c_1\|s_{k-1}\|^2$ for some positive constant $c_1 > 0$, it could be verified that $\|E_k\| \leq C\{\|B_{k-1} - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\|_2 + \|\widehat{\theta}_{\text{stage},k-1} - \widehat{\theta}_{\text{ge}}\|\}$ with probability tending to 1. This finishes the proof of the second part.

Next, applying (A.4) and the inductive method again, it could be proved that $\|\widehat{\theta}_{\text{stage},k} - \widehat{\theta}_{\text{ge}}\| = O_p(p^k(\log p)^{(k+1)/2}/n^{k+1}) + o_p(1/\sqrt{N})$. As a consequence, under the condition $N\{p^{2k}(\log p)^{k+1}\}/n^{2k+2} \to 0$, we have $\|\widehat{\theta}_{\text{stage},k} - \widehat{\theta}_{\text{ge}}\| = o_p(N^{-1/2})$, which accomplishes the whole corollary proof.

Appendix B: Some Useful Lemmas

Lemma 1 Assume the technical conditions (C1)–(C6) hold. Then, the following equations hold for some positive constants C_1–C_4 and $1 \leq k \leq 4$.

$$E\{\|\widehat{\theta}_{(m)} - \theta_0\|^k\} \leq C_1 n^{-k/2} C_G^2\{1 + o(1)\} \tag{B.1}$$

$$E\{\|\ddot{\mathcal{L}}_{(m)}(\theta_0) - \Omega(\theta_0)\|_2^k\} \leq C_2 \frac{\log^{k/2}(2p) C_H^k}{n^{k/2}} \tag{B.2}$$

$$E[\|\widehat{\theta}_{\text{stage},0} - \theta_0\|^2] \leq \left\{\frac{2C_G^2}{\tau_{\min}^2 N} + \frac{C_3 C_G^2}{\tau_{\min}^4 n^2}\left(C_H^2 \log p + \frac{C_{\max}^2 C_G^2}{\tau_{\min}^2}\right)\right\}\{1 + o(1)\} \tag{B.3}$$

$$E\left\|\{\ddot{\mathcal{L}}_{(m)}(\theta_0) - \dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)} - \theta_0) \otimes I_p\}\right\|_2 \leq C_4 \frac{p\sqrt{\log p}}{n}\{1 + o(1)\} \tag{B.4}$$

Proof. Given (B.1)–(B.3) in Theorem 1, B.0.1, Lemma 7 in Zhang et al. (2013) and (C6), it suffices to verify (B.4). To this end, operator $G_{(m)} = \dot{\mathcal{L}}_{(m)} - \dot{E}(\mathcal{L}_{(m)})$, and $G_{(m)}^j$ represents the jth element of $G_{(m)}$; then we have $\{\ddot{\mathcal{L}}_{(m)}(\theta_0) - \dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)} - \theta_0) \otimes I_p\} = [\ddot{G}_{(m)}^1(\theta_0)(\widehat{\theta}_{(m)} - \theta_0), \ldots, \ddot{G}_{(m)}^p(\theta_0)(\widehat{\theta}_{(m)} - \theta_0)]$. Consequently,

by the Cauchy–Schwarz inequality, it could be proved that

$$E\left\|\left\{\ddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\right\}\left\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\right\}\right\|_2 \le \sum_{j=1}^{p} E\left\|\ddot{G}^j_{(m)}(\theta_0)(\widehat{\theta}_{(m)}-\theta_0)\right\|$$
$$\le \sum_{j=1}^{P}\left\{E\|\ddot{G}^j_{(m)}(\theta_0)\|_2^2 E\|\widehat{\theta}_{(m)}-\theta_0\|^2\right\}^{1/2}$$

The first inequality holds because $\|B\|_2 \le \|B\|_F$ for any matrix B, and $\|\cdot\|_F$ represents the Frobenius norm. By Lemma 16 in Zhang et al. (2013), we have $E\|\ddot{G}^j_{(m)}(\theta_0)\|_2^2 \le O(\log p/n)$. This leads to $E\|\{\ddot{\mathcal{L}}_{(m)}(\theta_0)-\dot{\Omega}(\theta_0)\}\{(\widehat{\theta}_{(m)}-\theta_0)\otimes I_p\}\|_2 \le O(p\sqrt{\log p}/n)$. □

Lemma 2 Assume the technical conditions (C1)–(C6) hold. Then, we have $P(\bigcup_{m=0}^{M}\mathcal{E}_m^c)\to 0$, where $\mathcal{E}_m$s are defined in (A.1).

Proof. The proof is shown in Lemma 7 in Zhang et al. (2013) and D.1 in Jordan et al. (2019). □

Lemma 3 Assume the technical conditions (C1)–(C6) hold. Let new events

$$\mathcal{E}'_{(m)} = \Bigg\{\lambda_{\min}\left\{\ddot{\mathcal{L}}(\theta)\right\} \ge \frac{(1-\delta)\tau_{\min}}{2} \text{ for } \theta\in\{\theta_0,\widehat{\theta}_{\text{stage},0},\widehat{\theta}_{\text{ge}}\},$$
$$\|\widehat{\theta}_{\text{stage},0}-\widehat{\theta}_{\text{ge}}\| \le \frac{\tau_{\min}}{2C_{\max}} := \delta',\ \lambda_{\min}\left\{\ddot{\mathcal{L}}_{(m)}(\theta)\right\} \ge \frac{(1-\delta)\tau_{\min}}{2} \text{ for } \theta\in\{\theta_0,\widehat{\theta}_{(m)}\},$$
$$\max_{\theta\in B(\hat{\theta}_{\text{ge}},\delta')}\|\ddot{\mathcal{L}}(\theta)\|_2 \le 2C_{\max}\delta' + \frac{\delta\tau_{\min}}{4} + \tau_{\max} := C'_{\max}\Bigg\}$$

We have $P(\bigcap_{m=0}^{M}\mathcal{E}_m) < P(\bigcap_{m=0}^{M}\mathcal{E}'_m)$, where $\mathcal{E}_m$s are defined in (A.1).

Proof. We analyse the three terms under the event $\bigcap_{m=0}^{M}\mathcal{E}_m$ separately. First, we prove that

$$\lambda_{\min}\left\{\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})\right\} \ge \lambda_{\min}\left\{\Omega(\theta_0)\right\} - \|\ddot{\mathcal{L}}(\theta_0)-\Omega(\theta_0)\|_2 - \|\ddot{\mathcal{L}}(\widehat{\theta}_{\text{stage},0})-\ddot{\mathcal{L}}(\theta_0)\|_2$$
$$\ge \tau_{\min} - \frac{\delta\tau_{\min}}{2} - 2C_{\max}\|\widehat{\theta}_{\text{stage},0}-\theta_0\|_2 \ge \frac{(1-\delta)\tau_{\min}}{2}$$

The first inequality holds because $\lambda_{\min}(B_1) = \min_{\|u\|=1} u^{\top}(B_1-B_2+B_2)u \ge \min_{u_1=1} u_1^{\top}(B_1-B_2)u_1 + \min_{u_2=1} u_2^{\top}B_2u_2 \ge -\|B_1-B_2\|_2 + \lambda_{\min}(B_2)$ for any symmetric matrixes B_1 and B_2. The last inequality holds because $\|\widehat{\theta}_{\text{stage},0}-\theta_0\| \le \tau_{\min}/(4C_{\max})$ under $\bigcap_{m=0}^{M}\mathcal{E}_m$. By similar technical analysis, we know that $\lambda_{\min}\{\ddot{\mathcal{L}}(\theta)\} \ge (1-\delta)\tau_{\min}/2$ when $\theta=\theta_0$ or $\theta=\widehat{\theta}_{\text{ge}}$ and $\lambda_{\min}\{\ddot{\mathcal{L}}_{(m)}(\theta)\} \ge (1-\delta)\tau_{\min}/2$ for $\theta\in\{\theta_0,\widehat{\theta}_{(m)}\}$. Next, it is obvious that $\|\widehat{\theta}_{\text{stage},0}-\widehat{\theta}_{\text{ge}}\| \le \|\widehat{\theta}_{\text{stage},0}-\theta_0\| + \|\widehat{\theta}_{\text{ge}}-\theta_0\| \le \tau_{\min}/(2C_{\max})$ under $\bigcap_{m=0}^{M}\mathcal{E}_m$. Moreover, using the triangle inequality, we have

$$\max_{\theta\in B(\hat{\theta}_{\text{ge}},\delta')} \|\ddot{\mathcal{L}}(\theta)\|_2 \le \max_{\theta\in B(\hat{\theta}_{\text{ge}},\delta')} \|\ddot{\mathcal{L}}(\theta)-\ddot{\mathcal{L}}(\theta_0)\|_2 + \|\ddot{\mathcal{L}}(\theta_0)-\Omega(\theta_0)\|_2 + \|\Omega(\theta_0)\|_2$$

$$\le 2C_{\max}\delta' + \frac{\delta\tau_{\min}}{4} + \tau_{\max}$$

This accomplishes the whole lemma proof. □

Lemma 4 Assume the technical conditions (C1)–(C6) hold. Then, there exists a positive constant $\rho_K < 1$ such that for any $K > 0$, $\|\widehat{\theta}_{\text{stage},K} - \widehat{\theta}_{\text{ge}}\| \le \rho_K^{K-1}\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|$ with probability tending to 1.

Proof. We first verify the lemma when $K = 2$. For the SR1 update,

$$\begin{aligned}
\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{ge}} &= \widehat{\theta}_{\text{stage},1} - H_1\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) - \widehat{\theta}_{\text{ge}} \\
&= \widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}} - H_1\{\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) - \dot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\} \\
&= \widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}} - H_1\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})(\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}) \\
&\quad + H_1\Big[\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\big(\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\big) - \big\{\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) - \dot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\big\}\Big] \\
&= H_1\big\{B_1 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\big\}\big(\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\big) + \mathcal{O}
\end{aligned}$$

Here $\mathcal{O} = H_1[\ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})(\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}) - \{\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},1}) - \dot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}]$. By Taylor's expansion, $\mathcal{O}$ is a negligible higher-order term. Then, it suffices to analyse $H_1\{B_1 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\}$. We have

$$B_1 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}}) = B_0 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}}) + \frac{(y_0 - B_0s_0)(y_0 - B_0s_0)^\top}{(y_0 - B_0s_0)^\top y_0}$$

By similar analysis to that in Appendix A, Sections A.1 and A.2, it could be easily found that $B_0 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})$ and $\{(y_0 - B_0s_0)^\top y_0\}^{-1}(y_0 - B_0s_0)(y_0 - B_0s_0)^\top$ both converge to 0 in probability. Consequently, with probability tending to 1, there exists a small positive number $\rho_1 < 1$, such that $\|B_1 - \ddot{\mathcal{L}}(\widehat{\theta}_{\text{ge}})\|_2\|H_1\|_2 \le \rho_1/2$. Thus, we prove that $\|\widehat{\theta}_{\text{stage},2} - \widehat{\theta}_{\text{ge}}\| \le \rho_1\|\widehat{\theta}_{\text{stage},1} - \widehat{\theta}_{\text{ge}}\|$.

Similarly, we obtain the linear convergence rate for $\widehat{\theta}_{\text{stage},K}$ using the same analytical techniques used to study $\widehat{\theta}_{\text{stage},2}$. This finishes the lemma for SR1 updating. The proof for BFGS updating is shown in Lemma 3 in Mokhtari et al. (2018). □

Appendix C: Additional Numerical Details

C.1 Distributed *K*-stage SR1 algorithm

We introduce the detailed multi-stage algorithm with the SR1 updating strategy, not specified in Section 2.4. The key idea of the SR1 updating strategy is the same as that of BFGS updating strategy; that is, we establish a distributed version from the classical single computer updating formula. Nevertheless, the denominator in (2.1) involves $H^{(t)}$. As a result, we need to design a distributed algorithm more skilfully, so that the updated matrix of the distributed version is equivalent to that of the global one, and the number of communication rounds remains the same as that of the distributed BFGS updating algorithm. We first define $v_k = \widehat{\theta}_{\text{stage},k+1} - \widehat{\theta}_{\text{stage},k} - H_{k-1}\{\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},k+1}) - \dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},k})\}$ for any $k \ge 1$. In particular, we denote $H_{(m,-1)} = H_{(m,0)}$ and $v_{-1} = y_{-1} = 0$. The specific algorithm is given in Algorithm 4.

C.2 Distributed *K*-stage Newton–Raphson algorithm

Next, we introduce the detailed multi-stage algorithm but using the Newton–Raphson updating strategy, which is not specified in Section 3.2.

Appendix D: Updating Method of Quasi-Newton Matrix

We introduce the detailed intuition and proofs to derive (2.1) and (2.2) in the main text. To this end, denote $y^{(t)} = \dot{\mathcal{L}}(\widehat{\theta}^{(t+1)}) - \dot{\mathcal{L}}(\widehat{\theta}^{(t)})$ and $s^{(t)} = \widehat{\theta}^{(t+1)} - \widehat{\theta}^{(t)}$.

1. SR1 Updates. Equation (2.1) is the simplest quasi-Newton matrix updating formula. Let $H^{(t)}$ be the tth approximated Hessian inverse; we then derive $H^{(t+1)}$, satisfying the secant condition in (1.2), using rank one updating. To this end, we use the undetermined coefficient method, assuming that

$$H^{(t+1)} = H^{(t)} + \alpha u u^{\top} \tag{D.1}$$

for some undetermined coefficient $u \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}$. Then, according to (1.2), we have $s^{(t)} = H^{(t+1)} y^{(t)} = (H^{(t)} + \alpha u u^{\top}) y^{(t)}$. Then, it could be proved that

$$\alpha u^{\top} y^{(t)} u = s^{(t)} - H^{(t)} y^{(t)} \tag{D.2}$$

Note that $\alpha u^{\top} y^{(t)} \in \mathbb{R}$ is a scale, indicating that u and $s^{(t)} - H^{(t)} y^{(t)}$ share the same direction. Hence, we denote $u = s^{(t)} - H^{(t)} y^{(t)}$; then (D.2) could be rewritten as $\alpha (s^{(t)} - H^{(t)} y^{(t)})^{\top} y^{(t)} (s^{(t)} - H^{(t)} y^{(t)}) = s^{(t)} - H^{(t)} y^{(t)}$. Thus, we have $\alpha = \{(s^{(t)} - H^{(t)} y^{(t)})^{\top} y^{(t)}\}^{-1}$. Applying the results back to (D.1) leads to

$$H^{(t+1)} = H^{(t)} + \frac{\left(s^{(t)} - H^{(t)} y^{(t)}\right)\left(s^{(t)} - H^{(t)} y^{(t)}\right)^{\top}}{\left(s^{(t)} - H^{(t)} y^{(t)}\right)^{\top} y^{(t)}} \tag{D.3}$$

According to the Sherman–Morrison equation (Burden et al., 2015, Theorem 10.8), (D.3) could be rewritten as

$$B^{(t+1)} = B^{(t)} + \frac{\left(y^{(t)} - B^{(t)} s^{(t)}\right)\left(y^{(t)} - B^{(t)} s^{(t)}\right)^{\top}}{\left(y^{(t)} - B^{(t)} s^{(t)}\right)^{\top} s^{(t)}} \tag{D.4}$$

Here $B^{(t)} = \{H^{(t)}\}^{-1}$. When using the SR1 updating formula, it should be well defined. Consequently, (D.3) and (D.4) would be used only if

$$\begin{aligned} &\left|\left(s^{(t)} - H^{(t)} y^{(t)}\right)^{\top} s^{(t)}\right| \geq c_1 \|s^{(t)} - H^{(t)} y^{(t)}\| \|s^{(t)}\| \text{ or} \\ &\left|\left(y^{(t)} - B^{(t)} s^{(t)}\right)^{\top} s^{(t)}\right| \geq c_1 \|y^{(t)} - B^{(t)} s^{(t)}\| \|s^{(t)}\| \end{aligned} \tag{D.5}$$

for some positive constant $0 < c_1 < 1$. Otherwise, we keep $H^{(t+1)} = H^{(t)}$ or $B^{(t+1)} = B^{(t)}$; see more discussions in Conn et al. (1991) and Nocedal and Wright (1999).

2. SR2 (BFGS) Updates. SR1 updating is simple and easy to conduct. However, the positive definiteness of the approximated matrix (i.e. $H^{(t)}$) cannot be guaranteed; that is, we cannot ensure $(y^{(t)} - B^{(t)} s^{(t)})^{\top} s^{(t)} > 0$. The SR2 updating formula was proposed to address this problem. Similar to SR1 updating, given $H^{(t)}$, we consider the determined coefficient method to obtain the updating matrix $H^{(t+1)}$. Here, for simplicity, instead of directly analysing $H^{(t)}$, we consider $B^{(t)} = \{H^{(t)}\}^{-1}$ first. Thus, given $B^{(t)}$, assume that

$$B^{(t+1)} = B^{(t)} + a u u^{\top} + b v v^{\top}$$

Algorithm 4 Distributed K-stage quasi-Newton (SR1) algorithm

Input: DQN($K-1$) estimator $\widehat{\theta}_{\text{stage},K-1}$, v_{K-3} on the central computer, $\widehat{\theta}_{\text{stage},K-2}$, $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},K-2})$, y_{K-3}, and Hessian inverse approximation $H_{(m,K-3)}$ on the m-th worker

Output: DQN(K) estimator $\widehat{\theta}_{\text{stage},K}$

The central computer broadcasts $\widehat{\theta}_{\text{stage},K-1}$ and v_{K-3} to each worker

for $m = 1, 2, \ldots, M$ (distributedly) **do**

  Compute $\dot{\mathcal{L}}_{(m)}(\widehat{\theta}_{\text{stage},K-1})$ and transfer it to the central computer

  Update local Hessian inverse approximation by

  $H_{(m,K-2)} = H_{(m,K-3)} + \left[v_{K-3}^{\top} y_{K-3}\right]^{-1} v_{K-3} v_{K-3}^{\top}$

end

The central computer computes $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},K-1}) = M^{-1} \sum_{m=1}^{M} \dot{\mathcal{L}}_{(m)}(\widehat{\theta}_{\text{stage},K-1})$ and broadcasts it to each worker

for $m = 1, 2, \ldots, M$ *(distributedly)* **do**

  Compute $v_{(m,K-2)} = s_{K-2} - H_{(m,K-2)} y_{K-2}$ and transfer it to the central computer

  Calculate $H_{(m,K-2)} \dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},K-1})$ and transfer it to the central computer.

end

The central computer computes $v_{K-2} = M^{-1} \sum_{m=1}^{M} v_{(m,K-2)}$ and

$p_{K-1} = M^{-1} \sum_{m=1}^{M} H_{(m,K-2)} \dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},K-1}) + v_{K-2} v_{K-2}^{\top} \dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},K-1}) / (v_{K-2}^{\top} y_{K-2})$,

and obtains $\widehat{\theta}_{\text{stage},K} = \widehat{\theta}_{\text{stage},K-1} - p_{K-1}$.

Algorithm 5 Distributed K-stage Newton algorithm

Input: $K-1$-stage estimator $\widehat{\theta}_{\text{stage},K-1}$ on the central computer

Output: K-stage estimator $\widehat{\theta}_{\text{stage},K}$

The central computer broadcasts $\widehat{\theta}_{\text{stage},K-1}$ to each worker

for $m = 1, 2, \ldots, M$ (distributedly) **do**

  Compute $\dot{\mathcal{L}}_{(m)}(\widehat{\theta}_{\text{stage},K-1})$ and transfer it to the central computer

end

The central computer computes $\dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},K-1}) = M^{-1} \sum_{m=1}^{M} \dot{\mathcal{L}}_{(m)}(\widehat{\theta}_{\text{stage},K-1})$ and broadcasts it to each worker

for $m = 1, 2, \ldots, M$ *(distributedly)* **do**

  Compute $\left\{\ddot{\mathcal{L}}_{(m)}(\widehat{\theta}_{\text{stage},K-1})\right\}^{-1} \dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},K-1})$ and transfer it to the central computer

end

The central computer computes

$\widehat{\theta}_{\text{stage},K} = \widehat{\theta}_{\text{stage},K-1} - M^{-1} \sum_{m=1}^{M} \left\{\ddot{\mathcal{L}}_{(m)}(\widehat{\theta}_{\text{stage},K-1})\right\}^{-1} \dot{\mathcal{L}}(\widehat{\theta}_{\text{stage},K-1})$.

where $u, v \in \mathbb{R}^p$ and $a, b \in \mathbb{R}$. When

$$\left\{\dot{\mathcal{L}}(\widehat{\theta}^{(t+1)}) - \dot{\mathcal{L}}(\widehat{\theta}^{(t)})\right\} = B^{(t+1)}(\widehat{\theta}^{(t+1)} - \widehat{\theta}^{(t)}) \tag{D.6}$$

we obtain $(B^{(t)} + auu^{\top} + bvv^{\top}) s^{(t)} = y^{(t)}$; this leads to

$$\left(au^{\top} s^{(t)}\right) u + \left(bv^{\top} s^{(t)}\right) v = y^{(t)} - B^{(t)} s^{(t)}$$

Of the many ways to determine u and v, we consider the following criterion: $u = y^{(t)}$, $au^\top s^{(t)} = 1$, $v = B^{(t)}s^{(t)}$, and $bv^\top s^{(t)} = -1$. Consequently, (D.6) could be rewritten as

$$B^{(t+1)} = B^{(t)} + \frac{y^{(t)}\{y^{(t)}\}^\top}{\{s^{(t)}\}^\top y^{(t)}} - \frac{B^{(t)}s^{(t)}\{B^{(t)}s^{(t)}\}^\top}{\{s^{(t)}\}^\top B^{(t)}s^{(t)}}$$

Finally, according to the Sherman–Morrison formula (Burden et al., 2015, Theorem 10.8), we obtain the updating formula (2.2).

Moreover, there is another method to derive (2.2). To be more precise, $H^{(t+1)}$ is exactly the solution of the following optimal problem:

$$\begin{aligned} &\min_H \|H - H^{(t)}\|_W \\ &\text{s.t.} H = H^\top,\ Hy^{(t)} = s^{(t)} \end{aligned} \tag{D.7}$$

Here $\|H\|_W = \|W^{1/2}HW^{1/2}\|_F$ represents the weighted Frobenius norm, and W could be any matrix that satisfies $Ws^{(t)} = y^{(t)}$. Analysing the optimal problem using (D.7), we find that the solution of the problem (i.e. $H^{(t+1)}$) is a matrix H that is the closest to $H^{(t)}$. In addition, $H^{(t+1)}$ should be symmetric and satisfy the secant condition (D.6). For convex loss functions, it leads to $(s^{(t)})^\top y^{(t)} > 0$; thus, when $H^{(t)}$ is a positive definite, $H^{(t+1)}$ would also be positive definite; see more details in Nocedal and Wright (1999).

Appendix E: Supplementary Numerical Results

In this subsection, we provide the supplementary numerical results, which are not presented in Section 3.1. Specifically, to evaluate the performance of our proposed DQN method in Example 2, we report the log(MSE), SD, IQR, and range of log(MSE). The numerical results with different dimension p and local sample size n are given in Tables E6 and E7, respectively. All the results are qualitatively similar to those in the main text.

Table E6. Log(MSE) values and corresponding SD, IQR, and range for Examples 2

		Stage 0		Stage 1		Stage 2		Stage 3		Stage 4		
	p	SR1	BFGS	SR1	BFGS	SR1	BFGS	SR1	BFGS	SR1	BFGS	MLE
Log	1	−9.19	−9.19	−9.19	−9.19	−9.19	−9.19	−9.19	−9.19	−9.19	−9.19	−9.19
(MSE)	10	−6.84	−6.84	−6.91	−6.91	−6.91	−6.91	−6.91	−6.91	−6.91	−6.91	−6.91
	20	−6.09	−6.09	−6.20	−6.20	−6.21	−6.21	−6.22	−6.22	−6.22	−6.22	−6.22
SD	1	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
	10	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
	20	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
IQR	1	0.18	0.18	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19
	10	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
	20	0.05	0.05	0.03	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Range	1	0.66	0.66	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65
	10	0.30	0.30	0.26	0.26	0.27	0.27	0.26	0.26	0.27	0.27	0.27
	20	0.18	0.18	0.18	0.18	0.19	0.19	0.19	0.19	0.19	0.19	0.19

Note. The numerical performance is evaluated for different methods with different feature dimensions $p(\times 10^2)$. The whole sample size N and local sample size n are fixed to be 10^6 and 2×10^4, respectively. The reported results are averaged for $R = 100$ simulation replications.
MSE = mean squared error; SD = standard deviation; IQR = inter-quartile range; CR = coverage rate; BFGS = Broyden–Fletcher–Goldfarb–Shanno; MLE = maximum likelihood estimator.

Table E7. Log(MSE) values and corresponding SD, IQR, and range for Example 2

		Stage 0		Stage 1		Stage 2		Stage 3		Stage 4		
	n	SR1	BFGS	SR1	BFGS	SR1	BFGS	SR1	BFGS	SR1	BFGS	MLE
Log	50	−7.30	−7.30	−7.58	−7.58	−7.60	−7.60	−7.61	−7.61	−7.61	−7.61	−7.61
(MSE)	100	−7.49	−7.49	−7.60	−7.60	−7.60	−7.60	−7.61	−7.61	−7.61	−7.61	−7.61
	500	−7.60	−7.60	−7.60	−7.60	−7.61	−7.61	−7.61	−7.61	−7.61	−7.61	−7.61
SD	50	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
	100	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
	500	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
IQR	50	0.07	0.07	0.07	0.07	0.07	0.07	0.08	0.08	0.08	0.08	0.08
	100	0.07	0.07	0.07	0.07	0.07	0.07	0.08	0.08	0.08	0.08	0.08
	500	0.08	0.08	0.07	0.07	0.07	0.07	0.08	0.08	0.08	0.08	0.08
Range	50	0.39	0.39	0.36	0.36	0.37	0.37	0.37	0.37	0.37	0.37	0.37
	100	0.42	0.42	0.38	0.38	0.37	0.37	0.37	0.37	0.37	0.37	0.37
	500	0.39	0.39	0.36	0.36	0.36	0.36	0.37	0.37	0.37	0.37	0.37

Note. The numerical performance is evaluated for different $n(\times 10^2)$ and methods. The whole sample size N and feature dimension p are fixed to $N = 10^6$ and $p = 10^3$, respectively. Finally, the reported results are averaged based on $R = 100$ simulations.
MSE = mean squared error; SD = standard deviation; IQR = inter-quartile range; CR = coverage rate; BFGS = Broyden–Fletcher–Goldfarb–Shanno; MLE = maximum likelihood estimator.

References

Broyden C. G., Dennis Jr J. E., & Moré J. J. (1973). On the local and superlinear convergence of quasi-Newton methods. *IMA Journal of Applied Mathematics*, *12*(3), 223–245. https://doi.org/10.1093/imamat/12.3.223

Bubeck S. (2015). Theory of convex optimization for machine learning. *Foundations and Trends in Machine Learning*, *8*(3–4), 231–357. https://doi.org/10.1561/2200000050

Burden R. L., Faires J. D., & Burden A. M. (2015). *Numerical analysis*. Cengage Learning.

Chen W., Wang Z., & Zhou J. (2014). Large-scale L-BFGS using MapReduce. *Advances in Neural Information Processing Systems*, *27*. https://proceedings.neurips.cc/paper_files/paper/2014/file/e49b8b4053df9505e1f48c3a701c0682-Paper.pdf

Conn A. R., Gould N. I., & Toint P. L. (1991). Convergence of quasi-Newton matrices generated by the symmetric rank one update. *Mathematical Programming*, *50*(1–3), 177–195. https://doi.org/10.1007/BF01594934

Crane R., & Roosta F. (2019). DINGO: Distributed Newton-type method for gradient-norm optimization. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper_files/paper/2019/file/9718db12cae6be37f7349779007ee589-Paper.pdf

Davidon W. C. (1991). Variable metric method for minimization. *SIAM Journal on Optimization*, *1*(1), 1–17. https://doi.org/10.1137/0801001

Eisen M., Mokhtari A., & Ribeiro A. (2017). Decentralized quasi-Newton methods. *IEEE Transactions on Signal Processing*, *65*(10), 2613–2628. https://doi.org/10.1109/TSP.2017.2666776

Fan J., & Li R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Fan J., & Lv J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B*, *70*(5), 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

Fan J., & Song R. (2010). Sure independent screening in generalized linear models with NP-dimensionality. *Annals of Statistics*, *38*(6), 3567–3604. https://doi.org/10.1214/10-AOS798

Fan J., Wang D., Wang K., & Zhu Z. (2019). Distributed estimation of principal eigenspaces. *Annals of Statistics*, *47*(6), 3009. https://doi.org/10.1214/18-AOS1713

Goldfarb D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation*, *24*(109), 23–26. https://doi.org/10.1090/S0025-5718-1970-0258249-6

Gopal S., & Yang Y. (2013). Distributed training of large-scale logistic models. In *International conference on machine learning* (pp. 289–297). PMLR.

Goyal P., Dollár P., Girshick R., Noordhuis P., Wesolowski L., Kyrola A., Tulloch A., Jia Y., & He K. (2017). 'Accurate, large minibatch sgd: Training imagenet in 1 hr', arXiv, arXiv:1706.02677, preprint: not peer reviewed.

He X., Wang L., & Hong H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics*, *41*(1), 342–369. https://doi.org/10.1214/13-AOS1087

Hector E. C., & Song P. X.-K. (2020). Doubly distributed supervised learning and inference with high-dimensional correlated outcomes. *Journal of Machine Learning Research*, *21*(173), 173-1. https://doi.org/10.48550/arXiv.2007.08588

Hector E. C., & Song P. X.-K. (2021). A distributed and integrated method of moments for high-dimensional correlated data analysis. *Journal of the American Statistical Association*, *116*(534), 805–818. https://doi.org/10.1080/01621459.2020.1736082

Huang C., & Huo X. (2019). A distributed one-step estimator. *Mathematical Programming*, *174*(1–2), 41–76. https://doi.org/10.1007/s10107-019-01369-0

Jordan M. I., Lee J. D., & Yang Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, *114*(526), 668–681. https://doi.org/10.1080/01621459.2018.1429274

Lee C.-P., Lim C. H., & Wright S. J. (2018). A distributed quasi-Newton algorithm for empirical risk minimization with nonsmooth regularization. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1646–1655). https://doi.org/10.1145/3219819.3220075

Li G., Peng H., Zhang J., & Zhu L. (2012). Robust rank correlation based screening. *Annals of Statistics*, *40*(3), 1846–1877. https://doi.org/10.1214/12-AOS1024

Li X., Li R., Xia Z., & Xu C. (2020). Distributed feature screening via componentwise debiasing. *Journal of Machine Learning Research*, *21*(1), 1–32. https://dl.acm.org/doi/abs/10.5555/3455716.3455740

Lin S.-B., & Zhou D.-X. (2018). Distributed kernel-based gradient descent algorithms. *Constructive Approximation*, *47*(2), 249–276. https://doi.org/10.1007/s00365-017-9379-1

Mcdonald R., Mohri M., Silberman N., Walker D., & Mann G. S. (2009). Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in neural information processing systems* (pp. 1231–1239).

Mokhtari A., Eisen M., & Ribeiro A. (2018). IQN: An incremental quasi-Newton method with local superlinear convergence rate. *SIAM Journal on Optimization*, *28*(2), 1670–1698. https://doi.org/10.1137/17M1122943

Nocedal J., & Wright S. J. (1999). *Numerical optimization*. Springer.

Qu G., & Li N. (2019). Accelerated distributed Nesterov gradient descent. *IEEE Transactions on Automatic Control*, *65*(6), 2566–2581. https://doi.org/10.1109/TAC.2019.2937496

Schuller G. (1974). On the order of convergence of certain quasi-Newton-methods. *Numerische Mathematik*, *23*(2), 181–192. https://doi.org/10.1007/BF01459951

Shamir O., Srebro N., & Zhang T. (2014). Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning* (pp. 1000–1008). PMLR.

Shao J. (2003). *Mathematical statistics*. Springer texts in statistics. Springer.

Soori S., Mishchenko K., Mokhtari A., Dehnavi M. M., & Gurbuzbalaban M. (2020). DAve-QN: A distributed averaged quasi-Newton method with local superlinear convergence rate. In *Proceedings of the twenty third international conference on artificial intelligence and statistics* (pp. 1965–1976). PMLR.

Su L., & Xu J. (2019). Securing distributed gradient descent in high dimensional statistical learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, *3*(1), 1–41. https://doi.org/10.1145/3322205.3311083

Tang L., Zhou L., & Song P. X.-K. (2020). Distributed simultaneous inference in generalized linear models via confidence distribution. *Journal of Multivariate Analysis*, *176*, 104567. https://doi.org/10.1016/j.jmva.2019.104567

Van der Vaart A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge University Press.

Wang F., Zhu Y., Huang D., Qi H., & Wang H. (2021). Distributed one-step upgraded estimation for non-uniformly and non-randomly distributed data. *Computational Statistics & Data Analysis*, *162*, 107265.

Wang S., Roosta F., Xu P., & Mahoney M. W. (2018). Giant: Globally improved approximate newton method for distributed optimization. *Advances in Neural Information Processing Systems*, *31*, 2338–2348. https://dl.acm.org/doi/10.5555/3327144.3327160

Zhang Y., Duchi J. C., & Wainwright M. J. (2013). Communication-efficient algorithms for statistical optimization. *The Journal of Machine Learning Research*, *14*(1), 3321–3363. https://dl.acm.org/doi/10.5555/2567709.2567769

Zhang Y., & Lin X. (2015). DiSCO: Distributed optimization for self-concordant empirical loss. In *International conference on machine learning* (pp. 362–370). PMLR.

Zhu X., Li F., & Wang H. (2021). Least-square approximation for a distributed system. *Journal of Computational and Graphical Statistics*, *30*(4), 1–15. https://doi.org/10.1080/10618600.2021.1923517